The Dissertation Committee for DaLaine Chapman certifies that this is the approved version of the following dissertation:

Effects of Observation Duration on Evaluations of Teaching in Secondary School Band and Choir Rehearsals

Committee:

Robert A. Duke, Supervisor

Judith A. Jellison

Eugenia Costa-Giomi

Laurie P. Scott

Thomas J. O'Hare

Effects of Observation Duration on Evaluations of Teaching in Secondary School Band and Choir Rehearsals

by

DaLaine Chapman, B. Music Ed.; M. Music Ed.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2014

Effects of Observation Duration on Evaluations of Teaching in Secondary School Band and Choir Rehearsals

DaLaine Chapman, Ph.D.

The University of Texas at Austin, 2014

Supervisor: Robert A. Duke

The purpose of the present study was to determine whether expert evaluators' assessments of teachers vary between observations of rehearsal frames that demonstrate effective student behavior change and observations of full rehearsals. Ten experienced evaluators rated 12 music teachers on 10 criteria. The evaluators first observed brief video recordings of two rehearsal frames (RF) of each teacher and then a recording of a full rehearsal (FV) taught by the same teacher. The evaluators rated the teachers on all 10 criteria following each observation. Evaluators in the present study tended to rate teachers more highly and express greater confidence in their ratings in the FV condition than in the RF condition. These differences indicate that observing brief video episodes of teaching does not lead to the same ratings of teacher effectiveness as does observing video recordings of full rehearsals. The differences between the two conditions were larger in terms of evaluator confidence (29% higher confidence ratings in the FV condition) than in terms of ratings of teacher effectiveness (7% higher ratings in the FV condition). Although all teachers were rated more highly overall in the FV condition than in the RF condition, the differences between the two conditions were small and varied considerably among teachers and among evaluators.

Table of Contents

List of Tables

Chapter One: Introduction

Formal teacher evaluation is an increasingly prominent feature in professional practice. Because the success of students in school depends on quality teaching, a range of assessment procedures has been implemented nationwide ostensibly to ensure that children enrolled in public education receive high quality instruction (Danielson, 1996, 2001; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; Goldstein & Noguera, 2006; Goldstein, 2003, 2007; Marzano, 2007; Ovando & Harris, 1993; Ovando, 2001; Papay, 2012).

Largely as a result of national and state mandates, teachers in primary and secondary schools routinely undergo systematic evaluations, many of which include live observations of classroom instruction by administrators and other trained professionals (Danielson, 2001; Ovando & Ramirez, 2007; Peterson, 2004a). The process by which teachers are evaluated varies from state to state, but most often comprises a pre-observation conference, one or two full class observations, and a post-observation conference (Clements-Cortès, 2011; Danielson, 2001). These evaluations typically lead to formal assessment reports delivered to the teachers by the evaluators.

Most current evaluation systems are designed not only to document the competence of teaching faculty, but also to provide individual feedback that may be used to improve teachers' skills. Yet it remains to be determined whether the feedback conveyed in formal teacher evaluations contributes to increasing the teachers' effectiveness (Croft et al., 2011; Danielson, 2001; Darling-Hammond et al., 2012). If one of the functions of evaluative feedback is to refine the skills of teachers, then the feedback must be not only accurate but also meaningful to the teachers for whom it is intended.

QUESTIONS SURROUNDING EVALUATIONS

Efforts to optimize the quality and effectiveness of teacher evaluation raise questions about the nature of the assessment procedures employed and the means of conveying their results. Current debates surrounding the improvement of teaching assessments often center on the following questions: *Who should conduct evaluations? What should be evaluated? What should be the frequency and duration of teacher evaluations?* and *What should be the consequences of the results of teacher evaluations?*

Who should conduct teacher evaluations?

Typically, principals and assistant principals serve as the primary evaluators of teachers in public schools (Danielson, 2001; Darling-Hammond et al., 2012; Ovando, 2001; Papay, 2012; Peterson, 2004a). Although there may be cases when district personnel assist principals and assistant principals, most often it is the responsibility of the school-level administrators to complete teacher evaluations.

Of course, school evaluators seldom have specific expertise in all of the subject areas taught by the teachers they are charged to assess. This raises the question of whether they can accurately assess the work of teachers in the various disciplines taught in school. The notion of "pedagogical content knowledge" has gained currency in recent years (Mishra & Koehler, 2006; Veal & MaKinster, 1999), and it may be that aspects of teaching that are unique to individual disciplines may be beyond the purview of evaluators who lack expertise in those disciplines. Others have argued that there are fundamental elements of teaching that are generalizable across disciplines (Stein & Nelson, 2003), and thus individuals with expertise in teaching are fully capable of evaluating the effectiveness of teachers irrespective of the subject matter taught.

What should be evaluated?

Evaluators need to know where to look and what to assess when they observe teachers at work. Often evaluators use evaluation tools that comprise many indicators of effective teaching that have been adopted by state and local boards of education. These evaluation forms often define which aspects of teaching should be examined and where evaluators should focus their attention (Danielson & McGreal, 2000; Goldstein, 2003, 2004).

In addition, when searching for answers related to teachers' contributions to student growth over time, many states are now evaluating teachers on the basis of student standardized test scores, using value-added measures to mathematically account for how much of the students' progress may be attributed to a given teacher (Amrein-Beardsley & Collins, 2012; Koretz, 2008; Lockwood et al., 2007).

What should be the frequency and duration of teacher evaluations?

The most widely accepted practice is to evaluate teachers twice yearly; once during a full class session during the first semester (formative observation) and once during the second semester (summative observation); however, schedules vary among states (Clements-Cortès, 2011; Darling-Hammond et al., 2012; Wise, Darling-Hammond, McLaughlin, Berstein, 1985), and the timing and frequency of evaluations vary with the teachers' experience and expertise. For example, some districts evaluate inexperienced teachers and teachers whose performance in the past has been deemed unsatisfactory more frequently than they evaluate more experienced and successful teachers.

In addition to observations of full class periods, some districts include brief, random classroom observations that are completed several times per year. These

intermittent observations, referred to as "classroom walkthroughs," provide a snapshot of a teachers' work. These evaluations typically last only 5 to 7 minutes and are often unannounced. Because of the unplanned nature of a walkthrough, an evaluator may or may not see something that is informative in terms of teacher behavior.

What should be the consequences of the results of teacher evaluations?

The goals of teacher evaluations are not only to identify struggling teachers, but also to increase the effectiveness of teachers in every classroom so that every student has the opportunity to reach his potential; however, it may be difficult to establish credibility for a teacher evaluation process if the results have no meaningful implications for improving instruction. The possibility of linking evaluations to salary, tenure, remedial programs, and termination as a result of unsatisfactory evaluations currently exists in many states and districts.

TEACHER CHARACTERISTICS

Defining the attributes of good teaching across academic domains, including music, has been the subject of a vast amount of research (e.g., Coker, Medley, & Soar, 1980; Danielson, 1996, 2001; Darling-Hammond et al., 2012; Darling-Hammond, Wise, & Pease, 1983; Duke, 1999; Duke & Simmons, 2006; Heath & Nielson, 1974; Ovando & Ramirez, 2007; Ovando, 2001; Peterson, 2004a, 2004b; Scriven, 1988; Stronge, 2006). Identifying common characteristics of effective teaching can provide a framework for teacher evaluations that may ultimately result in formative evaluations that lead to staff development for those in need, and provide summative evaluations that are fair and equitable (Clements-Cortès, 2011).

Expert teachers engage in multiple behaviors during the course of a lesson, many occurring simultaneously. Writing about music instruction, Jellison (in press) states that highly-skilled teachers:

> determine which goals to teach and when; maintain students' attention; give individualized contingent feedback to groups and individuals; move efficiently and effectively through a carefully designed sequence of instruction; model and prompt when appropriate; teach for transfer of students' knowledge and skills; assess when learning has occurred; and throughout, demonstrate good musicianship, tenacity, sensitivity, appropriate seriousness and humor along the way. (p. 26)

Most would agree that the behaviors listed above are necessary for quality teaching. In fact, the majority of systematic observation forms currently in use include many of these same attributes.

Evaluation forms used in many observations have a multitude of individual components (sometimes as many as 60) for evaluators to mark when attempting to determine teachers' effectiveness; however, checklists of the individual components of effective teaching are not necessarily reliable predictors of successful learning in the classroom (Coker et al., 1980). It is the *effective combination* of the components, applied at the appropriate time and in the appropriate contexts, that lead to changes in student behavior (Coker et al., 1980; Medley & Coker, 1987; Stodolsky, 1984). Ultimately, being able to recognize and articulate the interdependencies among the components of effective teaching is necessary for evaluators to provide accurate and meaningful feedback that leads to improvements in teaching.

TIME ALLOTTED FOR EVALUATIONS

Few guidelines exist for how best to manage the time available for the evaluation process. The time available understandably holds important practical implications for

primary and secondary school teacher evaluation. The increasing demands associated with administrative positions often limit the time available for conducting effective evaluations (Hallinger & Heck, 1996; Keller, 1998; Krajewski, 1978; Ovando & Ramirez, 2007).

In a school that serves 450 students, for example, there may be two administrators on campus who conduct evaluations, leaving each evaluator with approximately 20 teachers to observe. With 20 teachers to evaluate, there may not be enough time to provide clear and meaningful feedback to each one, especially if there are teachers in need of specific remediation (Colby, Bradshaw, & Joyner, 2002; Danielson, 2001; Marshall, 2005; Stodolsky, 1984).

THIN SLICES AND DECISION-MAKING

Making quick decisions on the basis of what is often described as intuition or "gut feelings" is quite common. Because intuitive decisions often are made absent conscious awareness (Hogarth, 2001; Kahneman, 2002), it may be difficult for individuals charged with formal evaluations to explain precisely the basis of their assessment decisions. Human perceptions are based on a multitude of sensory data that interacts in complex and subtle ways with experiences stored in memory. Thus, intuitive judgments may be in many ways inexplicable (Ambady, 2010; Gigerenzer, 2007; Kahneman, 2013).

Evaluators form impressions about teachers; students form impressions about teachers; parents form impressions about their children's teachers; and teachers do the same with their peers. It is interesting that many of these judgments are fairly accurate, even though those making the judgements may be unable to fully explain how they came about. Quick decisions are typically made with a degree of automaticity, with the mind

combining environmental cues in ways that are below conscious awareness (Hogarth, 2001; Kahneman, 2013; Shapiro & Spence, 1997; Shirley & Langan-Fox, 1996).

Research indicates that *thin-slice* judgments (judgments based on limited time intervals) can sometimes be surprisingly accurate and may have a higher rate of accuracy than a carefully thought-out plan. Ambady and Rosenthal (1992) found in a meta-analysis of 38 experimental results an unexpectedly high rate of predictive accuracy when observers made judgments based on brief observations. Further research has continued to investigate the reliability of thin-slice judgments across various modes of input (e.g., visual, audio), obtaining similar results (Ambady & Rosenthal, 1992; Ambady, 2010).

Judgments about thin slices of behavior often capture a great deal of information about teacher effectiveness (Ambady & Rosenthal, 1993), and teacher expectations (teachers' expectations of students) (Babad et al., 1991). In fact, analyzing information before making a decision may actually disrupt intuitive judgments. Ambady (2010) examined whether distracters in cognitive processing disrupt or impair intuitive, automatic judgments. Thirty participants watched 39 10-second video clips of college teachers (see Ambady & Rosenthal, 1993). Participants were randomly assigned to one of four conditions: control, cognitive load, reasons-and-analysis, and delayed control. In the control condition, participants watched each clip, then provided a rating; in the cognitive-load condition, participants counted backwards from 1000 by 9's aloud while watching the clip, then provided a rating; in the reasons-and-analysis condition, participants watched each clip, then were asked to take one minute to write the rationale for their decision before recording their rating; and in the delayed control condition, the participants waited silently for 1 minute after watching the video before providing their rating. Results from the participants in each group were compared to the end-of-semester teacher evaluation ratings of the students who were in the teachers' classes for one

semester. The comparisons showed that the control, cognitive-load, and delayed-control groups were more closely correlated with the students' ratings; however, there was little correlation between the student's ratings of the teachers at the end of one semester and the participants' ratings in the reasons-and-analysis group. The longer the group spent analyzing *why* they were making the judgments of a teacher on the video clips, the more the analysis interfered with their more accurate, intuitive judgment. This finding suggests that quick decisions based on brief doses of information may in some circumstances be more accurate than carefully reasoned analyses.

REDUCING THE VARIABLES THAT EVALUATORS OBSERVE

When teachers are observed, evaluators may become overloaded by the number of variables they are supposed to assess and as a result, may miss important aspects of the teachers' behavior. Not all moments in a class or rehearsal are equally informative, of course. In fact, the quality of teaching varies from moment to moment in every classroom every day. Teachers produce instances of high information intermittently throughout a class period, and it may be that effective teachers evidence more of these high-information intervals than do their less effective colleagues. Instances that are devoted to bringing about changes in student behavior in the moment provide more information than do other instances when students engage in ongoing activity that elicits no behavior change that would be discernible to an observer. In music, a subject in which students are engaged in observable activity nearly all the time, there are potentially many such instances. An observer may see a teacher changing an embouchure of a wind player, correcting a bow hold of a string student, or addressing vowel placements in a choral class, all of which are readily observable in the moment.

It may be the case that these high-information intervals that reveal evidence of successful behavior change afford targets of observation on which to focus efforts in observation and evaluation. It may also be that these instances best exemplify the aspects of teaching that most clearly differentiate levels of teaching effectiveness. Is it possible to increase the incisiveness and efficiency of music teacher evaluation by focusing on intervals of instructional time?

To facilitate the observation of teaching and learning in music, Duke (1994) devised an approach to assessment that focuses on intervals of instructional time that are devoted to identifiable proximal learning goals, which he labeled rehearsal frames. This way of observing teaching focuses on the extent to which teachers bring about changes in student performance in the moment. Rehearsal frames have been applied in observing error correction in band rehearsals (Cavitt, 2003), expert wind conductors (Worthy, 2006), elementary music teachers (Taylor, 2006), choral teachers (Derby, 2001), conductors' rehearsal achievement (Montemayor, 2014), and Suzuki string teachers (Colprit, 2000).

PURPOSE OF THE STUDY AND RESEARCH QUESTIONS

Although teacher evaluation has been studied extensively, little attention has been devoted to determining the amount of time required to formulate reliable assessments of teachers' effectiveness. Using rehearsal frames as units of evaluation may provide a framework for assessing teachers' effectiveness efficiently and in ways that not only facilitate the work of evaluators but also provide more meaningful feedback to teachers.

I investigated whether experts' evaluations of selected rehearsal frames, ones that are high in information content as evidenced by clearly defined and accomplished targets, resemble the same experts' evaluations of full-length classes and rehearsals.

I compared experts' evaluations of brief and extended recorded episodes of instrumental and choral music teaching to determine whether their assessments of 10 dimensions of teacher effectiveness differed between the observation conditions. I compared the perceptions of 10 experienced evaluators who viewed two rehearsal frames and one full-rehearsal video of each of 12 music teachers. I posed the following questions:

1. To what extent are evaluators' assessments of teaching affected by the duration of the teaching episodes they observe? Do ratings of teaching effectiveness differ between observations of brief, targeted excerpts and observations of full class periods?
2. To what extent do evaluators' levels of confidence in their assessments differ between these two observation conditions?

The findings from this study may provide a basis for refining music teacher evaluation by creating a manageable context for observing the individual indicators of teaching effectiveness in tightly focused observations. In light of the challenges administrators face in balancing the demands of responsibilities other than observing and evaluating teachers, watching focused excerpts of teaching performance may provide a means of music teacher assessment that saves time and provides information that is useful in improving teaching. Such observations may also highlight the meaningful connections between what teachers do and what learners learn.

LIMITATIONS OF THE STUDY

All teacher participants in the present study were secondary band and choral teachers in public schools of a large district in Florida. The evaluator participants were 8 principals from the same Florida district and 2 Florida music supervisors. The 12 teacher participants taught in the same district as the principals, but evaluators did not rate teachers under their supervision.

None of the teacher participants were working under the supervision of any of the evaluators at the time of the study; however, given the number of years some of the teachers and evaluators had been concurrently employed in the district, it is perhaps not surprising that four of the evaluators had worked in the same schools as had four of the teachers; that is, each one of these four evaluators had worked in the same school along with one of the four teachers. In examining the evaluations, I found no indications that these former relationships produced any measurable effects on the ratings.

Evaluators in this district typically evaluate teachers during in-class observations, where both the teacher and students are in view. The video recordings for this project were focused on the teacher with very few of the students in view. The sound on the recordings, however, allowed evaluators to hear the student-teacher interactions and the students' performances clearly.

Evaluating teachers by observing videos of brief durations is clearly a departure from evaluators' typical experiences. It is certainly a departure from what the evaluators in this study had done in their home districts. These procedural differences should be considered carefully when interpreting the results.

Chapter Two: Review of Literature

There is a considerable body of research that focuses on teacher evaluation and assessment, including studies that examine attitudes and perception of evaluations (e.g., Kyriakides & Demetriou, 2007; Ovando, 2001; Ovando & Harris, 1993; Tuytens & Devos, 2010), collaboration between teachers and evaluators (Goldstein, 2003, 2007; Goldstein & Noguera, 2006), and standards-based and value-added measures (Glazerman et al., 2010; Palazuelos & Conley, 2008; Papay, 2012). Most of this research has examined teacher evaluation *within* specific school districts, but few studies have examined teacher evaluation data across a large number of districts.

Three extensive studies (Brandt, Mathers, Oliva, Brown-Sims, & Hess, 2007; Ellett & Garland, 1987; Loup, Garland, Ellett, & Rugutt, 1996) explored the applications of teacher evaluation procedures by collecting information from superintendents. In two of the three studies (Ellett & Garland, 1987; Loup, Garland, Ellett, & Rugutt, 1996) the authors analyzed evaluation information from the nation's 100 largest school districts; authors of the third study (Brandt, Mathers, Oliva, Brown-Sims, & Hess, 2007) described evaluation policies in a representative sample of districts in seven states: Illinois, Indiana, Iowa, Michigan, Minnesota, Ohio, and Wisconsin.

Ellett and Garland's (1987) survey yielded four important observations: (1) teacher evaluations emphasized summative (dismissal, remediation) rather than formative (professional development, teacher improvement) purposes; (2) most policies did not include requirements for establishing performance standards and evaluator training; (3) few districts permitted external or peer evaluations; and (4) superintendents seemed to be in favor of the policies set forth by their district. Additional results showed that principals

and assistant principals, rather than district personnel, were most likely to conduct the evaluations; and the frequency and duration of evaluations differed greatly among states.

Loup et al. (1996) conducted a follow-up study, also gathering information from superintendents of the nation's 100 largest school districts, to assess the potential impact of changing ideas about teacher evaluations since the research by Ellett and Garland. In the near decade between 1987 and 1996, teacher evaluation received increasing national attention with the development of new approaches for assessment; however, Loup et al. (1996) found few changes in teacher evaluation policies since Ellett and Garland's report; one change worth noting was that superintendents were more dissatisfied with their teacher evaluation procedures than those surveyed in Ellett and Garland's study and felt that their districts' existing evaluation procedures needed to be revised.

A decade later, Brandt et al. (2007) described teacher evaluation policies from a representative sample of districts in seven states—Illinois, Indiana, Iowa, Michigan, Minnesota, Ohio, and Wisconsin—and found that among staff and district personnel, principals and assistant principals most commonly conducted the evaluations, and only half of the districts specified when to evaluate teachers (e.g., Fall or Spring). Nearly all of the superintendents used the same evaluation form for all teachers, regardless of teacher experience levels, but the frequency of evaluations differed among teachers with different levels of experience. In five of the seven states, evaluation policies distinguished between beginning teachers, who were evaluated two or more times a year, and experienced teachers, who were evaluated once every two or three years.

Comparisons of results among the three studies indicate few policy changes in teacher evaluations had been implemented from the time of Ellett and Garland's work (1987) to that of Brandt et al. (2007).

SCHOOL REFORMS AND TEACHER EVALUATION

The history of federal school reform in the United States is punctuated by landmark efforts to change education policies. In 1965, the *Elementary and Secondary Education Act* (ESEA) (Elementary and Secondary Education Act of 1965) was passed as a part of President Lyndon Johnson's "War on Poverty." Central to ESEA's focus was the initiation of educational programs such as Title I and bilingual education. Since that time, several nationally led reports and reforms have provoked passion about the quality of American schools. In 1983, a report named *A Nation At Risk*, (National Commission on Excellence in Education, 1983) called for sweeping changes to public education and teacher training; and in 1994 President Bill Clinton attempted to bridge concerns about quality and equality of schools by promoting statewide standards and assessments through the *Improving America's Schools Act* (Improving America's Schools Act of 1994). Similar to the ESEA, this act included reforms in Title I, increased funding for bilingual and immigrant education, as well as provisions for public charter schools, drop-out prevention, and educational technology. In 1994 President Clinton also signed the *Goals 2000: Educate America Act* (Goals 2000: Educate America Act of 1994), which, as the names implies, challenged states to reach educational goals by the year 2000. By incorporating the lessons learned from previous educational reforms, this act gave states the freedom to establish aggressive reform plans that would be partially funded by the federal government.

Changes in educational policies were not limited to congressional reforms, as states were slowly beginning to restructure their own educational platforms. In 1993, Massachusetts passed the *Massachusetts Educational Reform Act* (Massachusetts Educational Reform Act of 1993), requiring common curriculum and statewide tests; and

as is often the case, other states followed Massachusetts' lead and implemented similar testing requirements.

In 2002, Congress amended the ESEA and reauthorized it as the *No Child Left Behind Act* (NCLB) (No Child Left Behind Act of 2002). Congress amended portions of NCLB in 2007, and in 2009 the Obama Administration introduced, *Race to the Top* (RTTT) (http://www2.ed.gov/).

The most recent changes in school reform began to focus on the central role of teachers in student achievement. In fact, one of NCLB's primary goals was that by 2014, every child would be taught by a highly qualified teacher.

Currently, in the majority of states, individual teacher performance is evaluated based in part on student achievement. According to a 2011 State of the States report issued by The National Council on Teacher Quality, there were 22 states along with the District of Columbia that evaluated teachers in part by students' standardized test scores; in the 2013 report, that number had risen to 40 (http://www.nctq.org/).

A few examples of how improving teacher effectiveness has taken center stage in an effort to improve student achievement include the Bill and Melinda Gates Foundation's (http://www.gatesfoundation.org/) *Measures of Effective Teaching* (MET) project (http://www.metproject.org/) which offers resources to districts to be used for increasing teacher effectiveness; the work of Charlotte Danielson and Robert Marzano, who introduced separate frameworks for effective teaching; states' efforts to assist teachers through the *Peer Assistance and Review* programs and the *Teacher Advancement Program*; and school districts across the United States who have developed formulas using so-called value-added measures in an attempt to determine how much of students' progress is attributable to individual teachers.

No Child Left Behind

In 2002, President George W. Bush signed into law the NCLB Act (No Child Left Behind Act of 2002). The primary goals of the act were to increase academic achievement and to close the achievement gap between White and minority students by focusing on three elements of policy: accountability, flexibility, and choice. Provisions of the act required that all public schools annually administer a statewide, standardized test to all students. In addition, all schools were required to make adequate yearly progress (AYP) in test scores. If scores were below what was expected and schools failed to meet AYP goals, free tutoring and other supplemental education services were offered to struggling students. If schools failed to meet AYP goals over multiple years, students enrolled in those schools were given the opportunity to transfer to higher performing schools. After the fourth year of failing to make AYP, an outside expert was appointed to advise the school; faculty and staff relevant to school failure were replaced; and school closings followed.

To accomplish the goals of NCLB, states were required to establish challenging academic standards for all schools and to test all students regularly to ensure that they were meeting those standards. NCLB also required that states and school districts hire teachers who were highly qualified in all schools, including Title I schools. Although states were permitted to set their own standards for determining which teachers were highly qualified, the law specified that to be qualified, teachers must have: a bachelor's degree, full state certification or licensure, and competence in the area of each subject they teach (http://www.ed.gov/). The goal of NCLB was that within 12 years of the act's passage (by 2014), all schools were to have all of their students scoring at the proficient level on state tests. *All* students included the various demographic subgroups enrolled in

schools, including those who were English-language learners and those receiving special education services.

The NCLB Act's implementation generated a great deal of controversy (Hoff, 2004). Proponents praised its goals and celebrated the rigorous accountability measures. Those who criticized the act questioned its emphasis on testing and claimed that it led teachers to teach to the test. When President Bush declared that he was in favor of "transforming the federal role in education so that no child is left behind" (NCLB, 2002, p. 1), people criticized the administration for interfering with state and local control over education while failing to fund all of the costs associated with the requirements of the act.

Race To The Top

RTTT is a school reform initiative that was authorized under the *American Recovery and Reinvestment Act* (ARRA) (American Recovery and Reinvestment Act of 2009), signed by President Barrack Obama. As the name implies, RTTT is a competition among states to secure a portion of $4.3 billion in federal funds for education. The initiative requires that states competing for funds fully adopt a set of common college and career-ready standards called the Common Core Standards, evaluate teachers based on gains in student achievement, emphasize content areas in science, technology, engineering, and mathematics (STEM), and restructure the lowest 5 percent of their schools (http://www.ed.gov/).

The RTTT initiative and the NCLB Act deal with many of the same issues and have many of the same goals, but their approaches are different; one provides incentives for schools to change, the other mandates it. The common goals are: high standards and

rigorous assessments, data collection and accountability, highly effective teachers and principals, and the restructuring of low-performing schools.

At the time of this writing, 36 states along with the District of Columbia have been awarded grants through five years of RTTT, but much like NCLB, RTTT has received mixed reviews. In a survey conducted by Harvard's Program on Education Policy and Governance, Peterson (2010) found that 32% of the general public supported RTTT, 22% opposed the initiative, and 46% had no opinion. There was a difference of opinion among teachers in the survey. When asked whether RTTT was "necessary to improve school quality" or whether it was an "unwarranted intrusion into state and local government," only 22% of teachers supported RTTT, while 46% of the teachers opposed the initiative.

The awarding of RTTT funds is in part influenced by states' demonstrating a strong commitment to the advancement of (STEM) education. Focusing on STEM subjects often involves schools' devoting a greater proportion of resources to these areas of study, which has led some critics to argue that such an emphasis results in less time available for the arts, humanities, languages, and physical education.

The Common Core State Standards Initiative

The Common Core State Standards Initiative (http://www.corestandards.org/) is a state-led effort that establishes a single set of educational standards for Kindergarten through 12th grade in English, language arts, and mathematics. State education leaders, assisted by teachers and school administrators, developed the Common Core Standards with the intent to provide a clear and consistent structure for students to successfully enter college and the workforce. The Common Core Standards Initiative is led by the

Nation's governors and education commissioners, through the National Governors Association (NGA) and the Council of Chief State School Officers (CCSSO).

The Common Core Standards were developed to provide teachers and parents with a common understanding of what students are expected to learn. Although the standards are not a curriculum, they are a clear set of goals and expectations for what knowledge and skills students will need to succeed. Teachers, principals, superintendents, and other district personnel decide how the standards are to be met. Since 2010, 45 states and the District of Columbia have adopted the Common Core Standards.

Much like RTTT and NCLB, the Common Core Standards Initiative has sparked controversy. Those who favor the initiative say that because it aligns goals and expectations from state to state, it will allow states to compare their student performance (Porter, McMaken, Hwang, & Yang, 2011). Supporters also argue that because of the increased rigor that the Common Core Standards are said to provide, students will be better prepared for college, a claim that remains to be verified.

Those who oppose the initiative state that the assessment tests currently do not include specific tests to accommodate students with special needs (modifications will not be in place for students with disabilities until the school year 2014-2015) (http://www.NCLD.org/), thus, all students in a school will have their results reported; technology and textbooks will need to be updated, involving large costs to school districts in order to satisfy the requirements. Those who argue against the initiative also say that it requires a difficult transition for students and teachers (Tienken & Canton, 2009). At the very least, teachers will have to challenge their students to think critically, which critics say is one of the most neglected areas of education. Ultimately, most teachers will have to prepare for how to develop critical thinking skills in their students.

TEACHER EVALUATION

National Board for Professional Teaching Standards

According to the National Board for Professional Teaching Standards (NBPTS) website (http://www.nbpts.org), since 1987, over 100,000 teachers have obtained National Board Teacher Certification. Not meant to replace a state teaching license, the national board certification is an advanced teaching credential for which teachers may voluntarily apply.

The NBPTS are based on five principles:

1. Teachers are committed to students and their learning

2. Teachers know the subjects they teach and how to teach those subjects to students

3. Teachers are responsible for managing and monitoring students' learning

4. Teachers think systematically about their practice and learn from experience

5. Teachers are members of learning communities

NBPTS candidates must complete 10 areas of assessments that are reviewed by a minimum of 12 trained teachers in the candidate's subject area. Included for submission are two components: four portfolio entries that show evidence of teaching practice, and exercises that assess content knowledge. Three of the portfolio entries are to be classroom-based, with video recordings and samples of student work serving as documentation of teaching quality. The fourth portfolio entry addresses relationships with the community and colleagues and demonstration of how these relationships impact student learning. Often requiring up to three years to complete, certification is granted for 10 years, and requires reapplication thereafter.

Measures of Effective Teaching

The Bill and Melinda Gates foundation devotes considerable resources to efforts aimed at improving the educational system in the United States. One way the foundation serves education is through the *College-Ready Education Program,* which ensures that students make successful transitions between high school and higher education and offers financial assistance by linking management consulting firms and technical assistance providers with the selected states to support the RTTT proposals. The Bill and Melinda Gates foundation also recognizes the financial pressure universities face with the demand of financial aid for low-income students. Therefore, another way the foundation serves education is by providing affordable access to post-secondary education intended to lead to a degree or certificate.

The foundation sponsors an educational initiative called *Measures of Effective Teaching* (MET) (http://www.metproject.org/), a research partnership among academics, teachers, and education organizations committed to investigating ways to identify and develop effective teaching. The goal of the MET project is to improve teacher effectiveness with information that will help districts build fair and reliable systems for teacher evaluation. This information can be used for a variety of purposes, including feedback and staff development, videotaped classroom observations, student surveys, tests of teachers' pedagogical content knowledge, and analyses of student assessment data to examine achievement gains over time.

Four reports have been published. The first report (December, 2010) focused on analyses of measures of student perceptions (student's rated their experiences with teachers in areas like caring and challenging lessons) and student achievement. The second report (January, 2012) gathered feedback by combining teacher observations with student surveys and information on achievement gains for students. The third and fourth

reports were released in 2013: one on the implications of assigning weights to different measures; another using random student assignments in classes to study the extent to which grouping students by ability may affect overall classroom test scores. It is an objective of the MET project to identify the effective teaching practices that ultimately improve student achievement.

Charlotte Danielson: The Danielson Group

Charlotte Danielson is the founder of the Danielson Group (http://www.danielsongroup.org/), which specializes in the design of teacher evaluation systems that promote professional learning and teacher improvement. Danielson has published multiple works that define effective teaching (Danielson, 2007), describe optimal structures for organizing schools (Danielson, 2001), and outline procedures for improving teacher leadership (Danielson, 2006). She is the creator of the *Charlotte Danielson Framework for Teaching,* which is a compilation of 22 components organized within four domains of instruction (a detailed review of the *Framework* appears later in this chapter). As an illustration of how researchers and educators are working together to improve classroom instruction, Domains 2 and 3 of Danielson's *Framework* have been incorporated in the Gates foundation's *Measures of Effective Teaching* (MET) project.

Robert Marzano: Marzano Evaluation Model

Robert Marzano is the co-founder and CEO of the Marzano Research Laboratory (http://www.marzanoresearch.com/) and is perhaps most widely known for his *Marzano Evaluation Model,* which includes 60 elements of teaching organized within four domains (a detailed review of the *Marzano Evaluation Model* appears later in this

chapter). Marzano has published multiple works on instruction (Marzano, Pickering, & Pollock, 2001), assessment (Marzano, Pickering, & McTighe, 1993), and supervision (Marzano, 1988).

Observation and evaluation software is also available with the *Marzano Teacher Evaluation Model*. iObservation is an instructional software tool that collects, manages, and reports data from observations conducted by an evaluator using a tablet or laptop. iObservation includes an extensive resource library that makes available video clips of effective teaching for teachers or evaluators to view, with specific targets of attention highlighted during a narration by Marzano. In addition, teachers can upload their own video clips to be viewed with an evaluator during a post observation conference.

Dal Lawrence: Peer Assistance and Review

Recent attempts to improve teacher evaluations have focused on guiding teachers toward the improvement of their practice. Districts nationwide are experimenting with different procedures for teacher evaluation, looking to replace what some say is a broken system of assessment in education (Papay, 2012). One such idea is a model of distributed leadership called *Peer Assistance and Review* (PAR). PAR departs from a traditional teacher evaluation in two important ways. First, master teachers are trained to conduct summative assessments (year-end assessments used for professional decisions like promotion, remediation, and salary) as well as formative assessments (assessments to provide information about strengths and weaknesses) of beginning teachers and veteran teachers in need of assistance (Goldstein, 2007). Second, PAR involves collaborations among teachers, peer teachers, administrators, and teachers' unions in efforts to improve the quality of instruction.

Created in 1981 by Dal Lawrence, a Toledo, Ohio, teachers' union president, PAR has been used in several states, including California, Ohio, and New York (Goldstein, 2007). With PAR, local teachers' unions and district administrators work together to improve teacher quality by having expert teachers mentor and evaluate their peers. Peer teachers complete extensive training and assist fellow teachers with lesson planning, classroom management, and implementing instruction.

In 1999, the California legislature began a statewide program that required all school districts to have a PAR model in place to serve veteran teachers receiving unsatisfactory evaluations. Whereas many districts had little formal structure in place to assist struggling veteran teachers, the PAR program was able to supply the support needed to improve the veterans' teaching.

In a longitudinal study of one California school district, Goldstein (2003b, 2004, 2007) found through interviews, observations, and surveys that, before PAR, teachers rarely had time for collaboration with their colleagues or time to reflect on their day-to-day activities. Since the implementation of PAR, frequent contacts between peer teachers and participating teachers led to assistance in planning and modeling lessons that led to overall instructional improvement.

PAR typically uses peer teachers, not principals, to serve as evaluators; however, Sullivan (2012) researched the history of the Montgomery (Maryland) school district as it attempted to formulate a teacher evaluation process that requires only first-year teachers to be evaluated by peer teachers, and veteran teachers to be evaluated by principals. Sullivan found that the principals' collaboration with the peer teachers in PAR was a major part of the success of the program.

As part of the program, principals and peer teachers undergo several hours of training to prepare them to conduct teacher evaluations. Ultimately, after a principal

makes her initial observation of a teacher and concludes that the teacher needs assistance, she enlists a peer teacher to perform separate visits. The visits by the peer teacher are to determine whether the teacher should enter the PAR program. For entry into the program, both principal and peer teacher must be in agreement. Once in the program, teachers have one year to improve, after which they are recommended for continued service or termination.

Lowell Milkin: Teacher Advancement Program

Another program that assists with the improvement of teachers is the *Teacher Advancement Program* (TAP). Created in 1999 by Lowell Milkin, TAP was developed as a system to attract, develop, motivate, and retain highly effective teachers (http://www.tapsystem.org/). TAP is based on four elements:

1. *Multiple Career Paths*, which, like PAR (Goldstein & Noguera, 2006), enlists skilled teachers to serve as master and mentor teachers.

2. *Ongoing Applied Professional Growth,* where teachers participate in weekly group meetings that are led by master teachers. The teachers examine student data, engage in collaborative planning, and observe master teachers modeling expert teaching.

3. *Instructionally Focused Accountability***,** where teachers are observed in classroom instruction several times per year by multiple trained observers.

4. *Performance-based Compensation***,** which offers bonuses each year based on teachers' demonstration of skills, knowledge, responsibilities, and their students' average growth in achievement. (http://www.niet.org/)

TAP is especially beneficial for schools serving high-need populations of students. A high-need school has been defined as a school where 30% or more of the students qualify for the federal free or reduced price lunch program due to low family income (http://www.ed.gov/). Historically, high-need schools have the most difficult time

staffing and retaining high-quality teachers from year to year (Jacob, 2007). A great deal of teacher turnover means "there are more new teachers every year to be mentored while at the same time there may be fewer highly skilled teachers available to provide peer support to teachers on campus" (Daley & Kim, 2010, p. 7).

Reliability and Validity

Reliability and validity define the quality of assessments in any domain. In the context of teacher evaluation, reliability refers to the "consistency of measurements across evaluators and observations" (Wise, Darling-Hammond, McLaughlin, Berstein, 1985, p. 89). Validity describes the extent to which the results of evaluations are accurate measures of teachers' actual effectiveness (Croft et al., 2011; Wise, Darling-Hammond, McLaughlin, Berstein, 1985). In order for the results of an evaluation to be valid, the evaluation process must align with its intended purpose. Yet, despite many years of application, teacher evaluation systems are often viewed by various constituencies (e.g., teachers, parents, community) as neither valid nor reliable (Noakes, 2009).

The task of designing effective systems of evaluation that can be universally applied across disciplines and grade levels is a challenging one, and issues of reliability and validity are deeply connected not only to the content of the evaluation instruments, but also to the backgrounds and experiences of evaluators and to the consequences associated with various evaluation outcomes. Typical teacher evaluation instruments that are based on observations of teachers' work specify numerous operationally-defined behaviors (or "indicators") in an effort to enhance the precision of the evaluations (Danielson, 2001; Kyriakides, 2005; Noakes, 2009), but the validity of the individual

criteria that teacher evaluations comprise are seldom assessed (Darling-Hammond et al., 1983; Kyriakides, 2005; Noakes, 2009; Peterson, 2004b).

The extent and complexity of the procedures employed in conducting observations for the purposes of evaluating teaching often require explicit evaluator training (Danielson, 2001; Wise, Darling-Hammond, McLaughlin, Berstein, 1985), although these observations are most often conducted by school principals and assistant principals (Danielson, 2001; Noakes, 2009; Ovando & Ramirez, 2007; Peterson, 2004). Local administrators may seem to be the most knowledgeable sources of information about teacher effectiveness, and thus the most appropriate individuals to conduct teacher evaluations, but questions remain about the requisite skills necessary to evaluate teaching effectively. Zimmerman and Deckert-Pelton (2003), for example, surveyed teachers in five Florida counties about their perceptions of their own principals' effectiveness as evaluators. Analyses of teachers' responses revealed four domains that defined effective evaluation: reciprocal, communicative interactions between evaluators (principals) and teachers; consistency both within and among schools in the implementation of evaluation procedures; teachers' perceptions of the principals' commitment; and the principals' apparent knowledge—particularly pedagogical content knowledge related to the disciplines begin evaluated—and experience.

Including measures of student learning in evaluations of teachers has become increasingly commonplace in recent years; however, identifying the role of teachers in relation to the other factors that contribute to student achievement, and doing so fairly and reliably, is difficult (Amrein-Beardsley & Collins, 2012). Value-Added Modeling (VAM) is often employed as a way to systematically quantify individual teachers' contributions to student learning. VAM uses as a dependent variable students' growth in a given academic year and considers multiple factors (one factor being the teacher) to

which students' growth may be attributed. In addition to prior test scores, value-added assessments include multiple variables in their formulas. Because of the many different variables that contribute to a students' growth, value-added formulas are often quite complex.

A considerable amount of research has examined whether VAM's fairly assess the contribution of teachers in terms of test score gains over time (e.g., Amrein-Beardsley & Collins, 2012; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; Koretz, 2008; Lockwood, McCaffrey, Hamilton, Stetcher, Le, & Martinez, 2007; Rothstein, 2009), and these studies have raised important questions about inherent biases in such measures. In terms of test score gains, Rothstein (2007) suggests that gains and losses shown by value-added measures may be biased because of the lack of randomization when assigning students to teachers. Schools often group students by "teams" sorted by academic success or failure. Rothstein (2009) concluded in another analysis that "even the best feasible value added models may be substantially biased, with the magnitude of the bias depending on the amount of information available for use in classroom assignments" (p. 1). Amrein-Beardsley and Collins (2012) point out that there is as yet no empirical evidence showing that employing value-added measures of teaching improves teachers' effectiveness or increases student achievement.

The results of teacher evaluations, in whatever form, may contribute to important decisions about teachers' financial compensation, employment, professional advancement, and recommendations for remediation. As such, the results of evaluations have meaningful consequences for teachers, students, and schools, which makes it particularly important that standards of reliability and validity of teacher evaluations remain high.

Methods of Observation

Formative and Summative Assessments

Efforts to increase school accountability have highlighted the importance of comprehensive evaluation procedures in identifying and maintaining quality teaching (e.g., Ovando & Harris, 1993; Ovando & Ramirez, 2007; Peterson, 2004). Both formative and summative evaluations are used on many school campuses and serve different roles in the overall evaluation process (Ovando & Ramirez, 2007). Whether assessments are labeled formative or summative depends primarily on the timing of the assessment. As the term suggests, formative assessments are typically conducted early in the year and provide information about areas of strength and weakness. Formative assessments are used as a basis for developing plans for improvement, and their results seldom lead to specific professional consequences. An assessment of a teacher is formative if its intent is to shape the behavior of the teacher (Garrison & Ehringhaus, 2007). Summative evaluations are year-end assessments that often are tied to decisions about salary, employment (e.g., retention, tenure), and professional status.

Range, Scherz, Holt, and Young (2011) state that the two processes of supervision (formative assessments) and evaluation (summative assessments) are outlined in much of the literature as two distinct techniques, and effective evaluators employ both to improve the quality of teaching. Zepeda (2006) contends that it is nearly impossible to separate the two forms of evaluation. Glickman, Cooper, and Ross-Cooper (2004) suggest that it is the responsibility of the evaluator to make certain that teachers understand the relationship between formative evaluations and teachers' professional growth. Thus, if evaluators do not make clear the procedure for improvement after the formative observation, teachers will continue to perceive the formative evaluation as a portion of the overall evaluation, instead of an opportunity for the improvement of instructional skills.

To provide sufficient information for meaningful evaluation, evaluators often conduct multiple observations throughout a school year, some of which are brief classroom visits called walk-throughs. Keruskin (2005) broadly described a walk-through as "an organized tour through the school using 'look-fors' to focus on instruction and learning" (p. 7). Look-fors are described as pre-defined student and teacher behaviors that are known to foster high student achievement. Classroom walk-throughs typically last from 5 to 7 minutes, yet during these brief visits it is possible for observers to take note of instructional practices, student engagement, classroom management, and other aspects of a teacher's work (Keruskin, 2005; Ovando & Ramirez, 2007).

Keruskin (2005) interviewed five high school principals over a four-year period and found that classroom walk-throughs positively affected teachers' self-efficacy, attitudes about professional development, teacher appraisal, classroom instruction, perceptions of principal effectiveness, and perceptions of school effectiveness. Data taken at the end of four consecutive years of using the classroom walk-through procedure showed "fewer students failing courses, fewer students repeating grade levels, an increase in SAT scores, and an increase in the graduation rate" (p.118).

Teachers whose administrators frequently visit their classrooms offering help through formative assessments may have an increased sense of self-efficacy. It is perhaps unsurprising that more frequent opportunities for evaluators to provide feedback to teachers may lead to improvements in teaching effectiveness.

Video Observations

Although live classroom observations are a more common feature of teacher evaluation than are evaluations of video recordings (e.g., Colby, Bradshaw, & Joyner,

2002; Danielson & McGreal, 2000; Darling-Hammond, Wise, & Pease, 1983; Haefele, 1993; Kimball & Milanowski, 2009), recent research has examined the use of video observations in providing feedback to novice teachers (e.g., Calandra, Brantley-Dias, Lee, & Fox, 2009) and induction teachers (those with fewer than five years' experience) (e.g., West, Rich, Shepherd, Recesso, & Hannafin, 2009). Other research has compared directly the results of video and live observations of pre-service teachers (e.g., Hartshorne, Heafner, & Petty, 2011) and experienced teachers (Casabianca et al., 2013).

Video recordings offer several potential advantages over live observations, providing enhanced focus on specific aspects of teaching (Calandra et al., 2009), affording opportunities for repeated viewings (Fagot & Hagan, 1988), and providing additional checks of reliability (Roberts & Hecht, 1996). Calandra et al. (2009) examined whether viewing edited video episodes of their own teaching would improve teachers' reflections about their work. Teachers who observed recordings of their teaching and identified meaningful instances of positive and negative teacher behavior produced more detailed written comments about their teaching than did others who did not view video recordings.

West et al. (2009) studied the ability of novice and experienced teachers to identify seven attributes of effective teaching, and found that observers' skills varied in relation to their experience levels. The observers found it easier to identify the target attributes than to rate the attributes along the evaluation continua provided. These data seem to illustrate that evaluating the quality of defined elements of teaching is more difficult than simply noting their presence.

One of the challenges involved in the observation and evaluation of teaching is the manifold nature of teaching episodes. Whether in music or in other disciplines, each class or rehearsal period comprises multiple components, and the individual intervals of

instructional time devoted to each component serve different purposes, and may focus on different aspects of the subject matter. This raises questions about whether there are intervals of instructional time that provide more information about teacher effectiveness than others, and whether focusing on intervals with the most information will increase efficiency and precision in assessment and evaluation. Duke (1994) suggested such a procedure for observing music instruction, arguing that intervals of instructional time that are devoted to identifiable proximal goals illustrate teachers' effectiveness in eliciting changes in student behavior.

These intervals, which Duke called rehearsal frames, may comprise any combination of teacher directives, modeling, questioning, and feedback, seen in relation to the behavior of students (Duke & Buckner, 2009). This narrowed observation focus has the potential to highlight the moment-to-moment changes in student behavior that may otherwise be overlooked during longer observations.

Teachscape

Founded in 1999, Teachscape is a web-based company that sells products and services that "bridge the gap between education research and everyday teaching practice" (http://www.teachscape.com). Teachscape programs have been implemented in school systems throughout the United States, including Washington, Alabama, Tennessee, and North Carolina. The company worked in partnership with the Bill and Melinda Gates Foundation, Educational Testing Service (ETS), and Charlotte Danielson to develop systematic teacher observation materials, including the *Framework for Teaching Proficiency System*. Recognizing that "the success of any classroom observation system ultimately depends on the accuracy and reliability of the individuals responsible for

observing and evaluating teachers" (http://www.teachscape.com), the developers of Teachscape designed a comprehensive system to help districts train evaluators that includes master-scored videos of effective teaching, observer and scorer training, and a proficiency test.

The procedures used to develop master-scored training videos are fascinating. During the school years 2009-2010 and 2010-2011, over 3000 teachers from Charlotte, North Carolina; Dallas, Texas; Denver, Colorado; Tampa, Florida; Memphis, Tennessee; and New York City, New York, provided videotapes of their teaching for inclusion into the Teachscape library. The Bill and Melinda Gates Foundation partnered with The Danielson Group to collect and analyze the data.

All classroom activity was recorded using panoramic digital video cameras located in participating teachers' classrooms. Participating teachers produced a written commentary about the teaching recording on the videos and uploaded both the videos and the commentaries to a secure Internet site. The commentary provided the observer with the context and background of the lesson as well as reflections of the lesson from the teacher. The videos were then watched and coded by independent observers drawn from a pool of teachers who teach similar subjects. The observers rated characteristics of teaching, such as the teachers' ability to provide useful feedback, explain concepts, manage student behavior, and create a positive learning environment. Approximately 1,500 videos were selected for inclusion into the Teachscape resource library to be used by districts that have implemented the Teachscape system.

Evaluation Models

Several models of teacher evaluation have become increasingly popular in recent years. Of the five states whose criteria I examined for potential use in this study, three of those states (Illinois, New York, and Florida) use adaptations of Danielson's *A Framework for Teaching*, or Marzano's *Marzano Teacher Evaluation Model*. The districts from California selected for examination in this study use three different evaluation models. Long Beach Unified School District uses the *California Standards for the Teaching Profession*, San Diego Unified School District uses an adaptation of the *California Standards for the Teaching Profession*; and the Los Angeles Unified School District uses the *Teaching and Learning Framework*. In Texas, several districts currently use the *Professional Development Appraisal System* (PDAS). An outline of the five states' evaluation criteria appear later in this chapter.

The 10 criteria that I chose for the evaluations conducted in this study appear in the *Marzano Teacher Evaluation Model*, the Danielson *A Framework for Teaching*, the *California Standards for the Teaching Profession*, and the *Professional Development Appraisal System* (Texas). The following is a summary of the *Marzano Teacher Evaluation Model*, and Danielson's *A Framework for Teaching*.

Marzano Teacher Evaluation Model

The *Marzano Teacher Evaluation Model* (http://www.marzanoresearch.com/) is purported to create causal links to raising student achievement (Marzano & Haystead, 2011). It is organized into 60 elements grouped under four domains. The 41 elements in Domain 1 describe teacher behaviors that are observable in the classrooms of effective

teachers, and are organized into 9 design questions (DQ's) grouped under three lesson segments that describe specific areas of instruction: *Lesson Segment Involving Routine Events, Lesson Segment Addressing Content,* and *Lesson Segment Enacted on the Spot*. The design questions are for teachers to use as a guide to track the progress of their teaching behavior. For example, the instructions for using the model suggest that teachers use the phrase, "What will I do to…." before each of the DQ's (e.g., "What will I do to communicate learning goals and provide feedback?").

Domain 1 focuses on in-class teacher behaviors. Domains 2, 3, and 4 pertain to teachers' planning and preparation, focusing on goal setting and decision making that foster high student achievement; reflection on teaching, which helps teachers evaluate their own instructional practices by using a professional growth plan; and collegiality and professionalism, which addresses the individual responsibility of teachers in promoting positive relationships with school and district colleagues. The elements of Domain 1 appear below.

Domain 1—Classroom Strategies and Behaviors

Lesson Segment Involving Routine Events

DQ 1: Communicating learning goals and feedback

1. Providing clear learning goals and scales
2. Tracking student progress
3. Celebrating student success

DQ 2: Establishing rules and procedures

4. Establishing classroom routines
5. Organizing the physical layout of the classroom

Lesson Segment Addressing Content

DQ 3: Helping students interact with new knowledge

6. Identifying critical information
7. Organizing students to interact with new knowledge
8. Previewing new content

9. Chunking content into " digestible bites"
10. Processing of new information
11. Elaborating on new information
12. Recording and representing knowledge
13. Reflecting on learning

DQ 4: Helping students to practice and deepen knowledge

14. Reviewing content
15. Organizing students to practice and deepen knowledge
16. Using homework
17. Examining similarities and differences
18. Examining errors in reasoning
19. Practicing skills, strategies, and processes
20. Revising knowledge

DQ 5: Helping students generate and test hypotheses

21. Organizing students for cognitively complex tasks
22. Engaging students in cognitively complex tasks involving hypothesis generation and testing
23. Providing resources and guidance

Lesson Segment Enacted on the Spot

DQ 6: Engaging Students

24. Noticing when students are not engaged
25. Using academic games
26. Managing response rates
27. Using physical movement
28. Maintaining a lively pace
29. Demonstrating intensity and enthusiasm
30. Using friendly controversy
31. Providing opportunities for students to talk about themselves
32. Presenting unusual or intriguing information

DQ 7: Recognizing adherence to rules and procedures

33. Demonstrating "with-it-ness"
34. Applying consequences for lack of adherence to rules and procedures
35. Acknowledging adherence to rules and procedures

DQ 8: Establishing and maintaining effective relationships with students

36. Understanding students' interests and backgrounds
37. Using verbal and nonverbal behaviors that indicate affection for students
38. Displaying objectivity and control

DQ 9: Communicating high expectations for all students

39. Demonstrating value and respect for low expectancy students
40. Asking questions of low expectancy students
41. Probing incorrect answers with low expectancy students

Danielson's A Framework for Teaching

Danielson's *A Framework for Teaching* (http://www.danielsongroup.org/) divides teacher behaviors into 22 components that are grouped under four domains. Similar to the Marzano model, *A Framework for Teaching* is designed to assess not only in-class teacher behavior but also other aspects of teachers' responsibilities, including planning and other professional work. As I did with the Marzano model, I have listed only those behaviors readily observable during in-class observations, which are Domains 1 through 3. The domains and their components appear below.

Domain 1 — Planning and Preparation

1. *Demonstrating knowledge of content and pedagogy*
    a. Knowledge of content and structure of the discipline
    b. Knowledge of prerequisite relationships
    c. Knowledge of content-related pedagogy

2. *Demonstrating knowledge of students*
    a. Knowledge of child and adolescent development
    b. Knowledge of the learning process
    c. Knowledge of students' skills, knowledge and language proficiency
    d. Knowledge of students' interests and cultural heritage
    e. Knowledge of students' special needs

3. *Setting instructional outcomes*
    a. Value, sequence, and alignment
    b. Clarity
    c. Balance
    d. Suitability for diverse learners

4. *Demonstrating knowledge of resources*
    a. Resources for classroom use
    b. Resources to extend content knowledge and pedagogy
    c. Resources for students

5. *Designing coherent instruction*
    a. Learning activities
    b. Instructional materials and resources
    c. Instructional groups
    d. Lesson and unit structure

6. *Designing student assessments*
    a. Congruence with instructional outcomes
    b. Criteria and standards
    c. Design of formative assessments
    d. Use for planning

Domain 2— Classroom Environment

1. *Creating an environment of respect and rapport*
    a. Teacher interaction with students
    b. Student interactions with one another

2. *Establishing a culture for learning*
    a. Importance of the content
    b. Expectations for learning and achievement
    c. Student pride in work

3. *Managing classroom procedures*
    a. Management of instructional groups
    b. Management of transitions
    c. Management of materials and supplies
    d. Performance of non-instructional duties
    e. Supervision of volunteers and paraprofessionals

4. *Managing student behavior*
    a. Expectations
    b. Monitoring of student behavior
    c. Responses to student misbehavior

5. *Organizing physical space*
    a. Safety and accessibility
    b. Arrangement of furniture and use of physical resources

Domain 3—Instruction

1. *Communicating with students*
    a. Expectations for learning
    b. Directions and procedures
    c. Explanations of content
    d. Use of oral and written language

2. *Using questioning and discussion techniques*
    a. Quality of questions
    b. Discussion techniques
    c. Student participation

3. *Engaging students in learning*
    a. Activities and assignments
    b. Grouping of students
    c. Instructional materials and resources
    d. Structure and pacing

4. *Using assessment in instruction*
    a. Assessment criteria
    b. Monitoring of student learning
    c. Feedback to students
    d. Student self-assessment and monitoring of progress

5. *Demonstrating flexibility and responsiveness*
    a. Lesson adjustment
    b. Response to students
    c. Persistence

State-Adopted Models

The following is an outline of in-class teacher evaluation criteria currently used in the three most populous districts in each of the five most populous states in the United States: California, Texas, Illinois, New York, and Florida.

California *(http://www.cde.ca.gov/)*

California's three largest districts—Los Angeles Unified School District (LAUSD), San Diego Unified School District (SDUSD), and Long Beach Unified School District (LBUSD)—use 3 different evaluation models to evaluate teachers. The school districts' models and criteria appear below.

The Los Angeles Unified School District (LAUSD) adopted the *Teaching and Learning Framework* (TLF), which lists 18 elements under 5 teaching standards. The LAUSD evaluates their teachers based on multiple measures of assessment, including artifacts (e.g., student work, lesson plans), and surveys (parent and student). The *Teaching and Learning Framework* Standards 4 (*Additional professional responsibilities*) and 5 (*Professional growth*) are excluded from this list as they address out-of-class behaviors. The three remaining standards are:

Standard 1 — Planning and preparation

1. Demonstrating knowledge of content and pedagogy
2. Demonstrating knowledge of students
3. Establishing instructional outcomes
4. Designing coherent instruction
5. Designing student assessment

Standard 2 — Classroom environment

1. Creating an environment of respect and rapport
2. Establishing a culture for learning
3. Managing classroom procedures
4. Managing student behavior

Standard 3 — Delivery of instruction

1. Communicating with students
2. Using questioning and discussion techniques
3. Structures to engage students in learning
4. Using assessment in instruction to advance student learning

The San Diego Unified School District teacher evaluation model defines 12 elements under 6 standards of teacher effectiveness. Standard 6 (*Teachers develop as professional educators*) is excluded from this list, as it comprises out-of-class teacher behaviors. The remaining five standards appear below.

Standard 1 — Teachers engage and support all students in learning

1. Differentiated instruction and teachers engage students in meaningful learning tasks that are relevant, authentic, and reflect real world situations

Standard 2 — Teachers create and maintain effective environments for student learning

1. Classroom observations
2. Use of effective behavioral strategies

Standard 3 — Teachers understand and organize subject matter

1. Employ varied instructional strategies
2. Demonstrate knowledge of content

Standard 4 — Teachers plan instruction and design learning experiences for all students

1. Use of student data in planning instruction
2. Use of effective research-based strategies for teaching English Language Learners (ELL) and special education students
3. Effective use of Response to Intervention (RTI) and Universal Design for Learning (UDL) strategies

Standard 5 — Teachers effectively assess student learning

1. Using frequent and formative assessments linked to state/federal expectations
2. Frequent monitoring of student data
3. Grading policies and grades reflect student learning
4. Facilitate high level complex conversations and discussions; Timely descriptive feedback to students

The Long Beach Unified School District evaluates their teachers using the *California Standards for the Teaching Profession,* which defines 32 elements under 6 domains of teaching standards. Standard 6 (*Developing as a professional educator*) is excluded from this list, as it comprises out-of-class teacher behaviors. The remaining standards appear below.

Standard 1— Engaging and supporting all students in learning

1. Connecting students' prior knowledge, life experience, and interests with learning goals

2. Using a variety of instructional strategies and resources to respond to students' diverse needs
3. Facilitating learning experiences that promote autonomy, interaction, and choice
4. Engaging students in problem solving, critical thinking, and other activities that make subject matter meaningful
5. Promoting self-directed, reflective learning for all students

Standard 2— Creating and maintaining effective environments for student learning

1. Creating a physical environment that engages all students
2. Establishing a climate that promotes fairness and respect
3. Promoting social development and group responsibility
4. Establishing and maintaining standards for student behavior
5. Planning and implementing classroom procedures and routines that support student learning
6. Using instructional time effectively

Standard 3— Understanding and organizing subject matter for student learning

1. Demonstrating knowledge of subject matter content and student development
2. Organizing curriculum to support student understanding of subject matter
3. Interrelating ideas and information within and across subject matter areas
4. Developing student understanding through instructional strategies that are appropriate to the subject matter
5. Using materials, resources, and technologies to make subject matter accessible to students

Standard 4—Planning instruction and designing learning experiences for all students

1. Drawing on and valuing students' backgrounds, interests, and developmental learning needs
2. Establishing and articulating goals for student learning
3. Developing and sequencing instructional activities and materials for student learning
4. Designing short-term and long-term plans to foster student learning
5. Modifying instructional plans to adjust for student needs

Standard 5— Assessing student learning

1. Establishing and communicating learning goals for all students
2. Collecting and using multiple sources of information to assess student learning
3. Involving and guiding all students in assessing their own learning
4. Using the results of assessments to guide instruction
5. Communicating with students, families, and other audiences about student progress

Texas *(http://www.tea.state.tx.us/)*

Several districts in the State of Texas, including the Dallas Independent School District and the Fort Worth Independent School District, currently use the *Professional Development Appraisal System* (PDAS). In 2011, the Houston Independent School District adopted the *Teacher Appraisal Development System*. Criteria from both models appear below.

The Professional Development Appraisal System (PDAS) includes 51 criteria within 8 domains. Two domains, VI (*Professional development*), and VII (*Compliance with policies, operating procedures, and requirements*) comprise out-of-class teaching behaviors and are not included on the following list:

Domain I — Active student participation in the learning process

1. Engaged in learning
2. Successful in learning
3. Critical thinking/problem solving
4. Self-directed
5. Connects learning

Domain II — Learner-centered instruction

1. Goals and objectives
2. Learner-centered
3. Critical thinking and problem solving
4. Motivational strategies
5. Alignment
6. Pacing/sequencing
7. Value and importance
8. Appropriate questioning and inquiry
9. Use of technology

Domain III — Evaluations and feedback on student progress

1. Monitored and assessed
2. Assessment and instruction are aligned
3. Appropriate assessment
4. Learning reinforced
5. Constructive feedback
6. Relearning and re-evaluation

Domain IV — Learning environment, time and materials

1. Discipline procedures
2. Self-discipline and self-directed learning
3. Equitable teacher/student interaction
4. Expectations for behavior
5. Redirects disruptive behavior
6. Reinforces desired behavior
7. Equitable and varied characteristics
8. Manages time and materials

Domain V — Professional communication

1. Written with students
2. Verbal/non-verbal with students
3. Reluctant students
4. Written with parents, staff, community members, and other professionals
5. Verbal/non-verbal with parents, staff, community members, and other professionals
6. Supportive, courteous

Domain VIII — Improvement of academic performance of all students on the campus

1. Aligns instruction
2. Analyzes *Texas Assessment of Knowledge and Skills* (TAKS) data
3. Appropriate sequence
4. Appropriate materials
5. Monitors student performance
6. Monitors attendance
7. Students in at-risk situations
8. Appropriate plans for intervention
9. Modifies and adapts

The Houston Independent School District (HISD) uses the *Teacher Appraisal Development System* which comprises three standards: (1) *Planning*, (2) *Instruction*, and (3) *Professionalism*. In addition, HISD evaluates their teachers based on students' performance on standardized tests. Standard 3 (*Professionalism*) is not listed as it addresses out-of-class behaviors. Standards 1 and 2 appear below.

Standard 1 — Planning

1. Develops student learning goals
2. Collects, tracks, and uses student data to drive instruction
3. Designs effective lesson plans, units, and assessments

Standard 2 — Instruction

1. Facilitates organized, student-centered, objective-driven lessons
2. Checks for student understanding and responds to student misunderstanding
3. Differentiates instruction for student needs by employing a variety of instructional strategies
4. Engages students in work that develops higher-level thinking skills
5. Maximizes instructional time
6. Communicates content and concepts to students
7. Promotes high academic expectations for students
8. Students actively participating in lesson activities
9. Sets and implements discipline management procedures
10. Builds a positive and respectful classroom environment

Illinois *(http://www.isbe.state.il.us/)*

Chicago, Elgin, and Rockford Public Schools currently have adopted *A Framework for Teaching* (see details of criteria on pages 37-39). Chicago uses the *Framework* in its entirety, while Elgin has excluded Domain 4 (*Professional responsibilities*). Rockford added a fifth domain (*Student achievement and growth*). In addition to *A Framework for Teaching*, all three school districts evaluate their teachers based on multiple measures of assessment. Chicago Public Schools includes student test scores and surveys of student feedback; Elgin Public Schools includes student artifacts (examples of student work), but does not include student test scores; and Rockford Public Schools include student test scores in the evaluation of their teachers.

New York *(http://www.nysed.gov/)*

New York City and Rochester Public Schools use *A Framework for Teaching* in its entirety (see details of criteria on pages 37-39) to evaluate their teachers. Buffalo Public Schools adopted the *New York State Teacher Practice Rubric*, which defines 36 elements that are grouped under 7 domains. In addition to each district's selected model (*A Framework for Teaching* or the *Teacher Practice Rubric*), the three school systems evaluate their teachers based on multiple measures of assessment. New York City uses student artifacts, student outcome data, and student feedback; Rochester and Buffalo use student standardized test scores as a component of a teacher's evaluation. *Teacher Practice Rubric* Domains 6 (*Professional responsibilities and collaboration*), and 7 (*Professional growth*) are excluded from this list as they address out-of-class teacher behaviors. The remaining domains and their components appear below:

Standard 1 — Knowledge of students and student learning

Goal: Teachers acquire knowledge of each student and demonstrate knowledge of student development and learning to promote achievement for all students

1.Teachers demonstrate knowledge of child and adolescent development, including students' cognitive, language, social, emotional, and physical developmental levels
2. Teachers demonstrate current, research-based knowledge of learning and language acquisition theories and processes
3. Teachers demonstrate knowledge of and are responsive to diverse learning needs, strengths, interests, and experiences of all students
4. Teachers acquire knowledge of individual students from students, families, guardians, and/or caregivers to enhance student learning
5. Teachers demonstrate knowledge of and are responsive to the economic, social, cultural, linguistic, family, and community factors that influence their students' learning
6. Teachers demonstrate knowledge and understanding of technological and information literacy and how they affect student learning

Standard 2 — Knowledge of content and instructional planning

Goal: Teachers know the content they are responsible for teaching and plan instruction that ensures growth and achievement for all students

1. Teachers demonstrate knowledge of the content they teach, including relationships among central concepts, tools of inquiry, structures and current developments within their discipline(s)
2. Teachers understand how to connect concepts across disciplines and engage learners in critical and innovative thinking and collaborative problem solving related to real world contexts
3. Teachers use a broad range of instructional strategies to make subject matter accessible
4. Teachers establish goals and expectations for all students that are aligned with learning standards and allow for multiple pathways to achievement
5. Teachers design relevant instruction that connects students' prior understanding and experiences to new knowledge
6. Teachers evaluate and utilize curricular materials and other appropriate resources to promote student success in meeting learning goals

Standard 3 — Instructional practice

Goal: Teachers implement instruction that engages and challenges all students to meet or exceed the learning standards

1. Teachers use research-based practices and evidence of student learning to provide developmentally appropriate and standards-driven instruction that motivates and engages students in learning
2. Teachers communicate clearly and accurately with students to maximize their understanding and learning
3. Teachers set high expectations and create challenging learning experiences for students
4. Teachers explore and use a variety of instructional approaches, resources, and technologies to meet diverse learning needs, engage students and promote achievement
5. Teachers engage students in the development of multi-disciplinary skills, such as communication, collaboration, critical thinking, and use of technology
6. Teachers monitor and assess student progress, seek and provide feedback, and adapt instruction to student needs

Standard 4 — Learning environment

Goal: Teachers work with all students to create a dynamic learning environment that supports achievement and growth

1. Teachers create a mutually respectful, safe, and supportive learning environment that is inclusive of every student
2. Teachers create an intellectually challenging and stimulating learning environment
3. Teachers manage the learning environment for the effective operation of the classroom
4. Teachers organize and utilize available resources (e.g. physical space, time, people, technology) to create a safe and productive learning environment

Standard 5 — Assessment for student learning

Goal: Teachers use multiple measures to assess and document student growth, evaluate instructional effectiveness, and modify instruction

1. Teachers design, select, and use a range of assessment tools and processes to measure and document student learning and growth
2. Teachers understand, analyze, interpret, and use assessment data to monitor student progress and to plan and differentiate instruction
3. Teachers communicate information about various components of the assessment system
4. Teachers reflect upon and evaluate the effectiveness of their comprehensive assessment system to make adjustments to it and plan instruction accordingly
5. Teachers prepare students to understand the format and directions of assessments used and the criteria by which the students will be evaluated

Florida *(http://www.fldoe.org)*

Florida's three largest school districts—Broward County Public Schools, Miami-Dade County Public Schools, and the School District of Hillsborough County—have adopted separate models from one another; however, all evaluate their teachers based on a Value-Added Modeling formula that is included in a teachers overall evaluation by Florida law. The districts' and their evaluation models appear below.

Broward County Public Schools adopted the complete *Marzano Teacher Evaluation Model* (see criteria on pages 35-36). In 2011, Broward began with Domain 1, and in the years following have since implemented all four domains. The School District

of Hillsborough County adopted *A Framework for Teaching* in its entirety (see criteria on pages 37-39).

Miami-Dade County Public Schools use what is called the *Individual Performance Evaluation and Growth System* (IPEGS) to evaluate their teachers. The system comprises seven standards of effective teaching based upon three foundational principles: a focus on high expectations, knowledge of subject matter, and the standards of the profession. Standard 6 is excluded from this list as it addresses out-of-class teacher behaviors. The remaining standards appear below:

Standard 1—Knowledge of learners

The teacher identifies and addresses the needs of learners by demonstrating respect for individual differences, cultures, backgrounds, and learning styles

Standard 2—Instructional planning

The teacher uses appropriate curricula (including state reading requirements, if applicable), instructional strategies, and resources to develop lesson plans that include goals and/or objectives, learning activities, assessment of student learning, and home learning in order to address the diverse needs of students

Standard 3—Instructional delivery and engagement

The teacher promotes learning by demonstrating accurate content knowledge and by addressing academic needs through a variety of appropriate instructional strategies and technologies that engage learners

Standard 4—Assessment

The teacher gathers, analyzes, and uses data (including *Florida Comprehensive Assessment Test* (FCAT) state data, as applicable) to measure learner progress, guide instruction, and provide timely feedback

Standard 5—Communication

The teacher communicates effectively with students, their parents or families, staff, and other members of the learning community

Standard 7—Learning environment

The teacher creates and maintains a safe learning environment while encouraging fairness, respect, and enthusiasm

CONCLUSIONS

During my tenure as a classroom teacher, principals and assistant principals evaluated my teaching using evaluation instruments that comprised checklists of teacher behaviors. When I became a music supervisor, I assisted school administrators with evaluations, often using the same evaluation procedures.

In the study reported in this document, I examined the effect of observation content and duration on evaluators' assessments of teaching. To develop a concise set of common criteria for use in this project, I examined the evaluation procedures and criteria used in the three most populous districts in each of the five most populous states in the United States (California, Texas, Illinois, New York, and Florida), looking for commonalities among them. There were many. A detailed review of the criteria, and how the criteria were chosen for this study appears in Chapter 3.

In the current debates surrounding educational assessment, questions remain about which aspects of teaching to examine and where evaluators should focus their attention (Danielson, 2007; Darling-Hammond et al., 2012; Goldstein, 2003a, 2004). Many districts and states are experimenting with evaluation practices that are linked to federal incentives. School reforms like NCLB and RTTT have influenced the ways that teachers are evaluated, and defining the most effective system of teacher evaluation remains an ongoing challenge.

Accurately identifying the characteristics of effective teaching, providing training for evaluators, and providing opportunities for meaningful formative and summative assessments are essential ingredients of successful teacher evaluation. Administrators and teachers alike assess the overall effectiveness of potential evaluation procedures in terms of both accuracy *and* efficiency as reflected in responses to programs developed to

improve teacher practice, like *Peer Assistance and Review* and the *Teacher Advancement Program* (Goldstein, 2007).

The criteria by which schools evaluate their teachers are similar among districts, yet for some school systems, the manner in which this information can be accessed remains unclear. In searching for common criteria for this study, most districts were transparent in their effort to provide detailed information about the evaluation procedures and criteria. Surprisingly, there were some districts that did not seem to have a highly systemized procedure in place, in which instances there was no mention of *RTTT* and very little mentioned about the Common Core Standards.

Searching through materials and navigating through different choices of language among districts (e.g., the terms standard, objective, benchmark, and strand used to describe similar concepts) highlights the need for consistency. The Common Core Standards are said to create consistency for what *students* should know and learn. While there is no single teacher evaluation procedure that has criteria for everything that teachers should know and do—one that fits the needs of all districts—the commercial models promoted by Marzano and Danielson seem to narrow the gap.

The criteria in districts that use commercially available models as a tool for teacher evaluation, of course, were easier to identify. Although the work of Danielson and Marzano has become well known throughout school systems nationwide, several districts in this study use other models that are comprehensive and data driven.

School reforms likely influenced recent changes to evaluation criteria and procedures, but again, not for all states. Texas, for example, elected not to compete for the RTTT funds, nor did it adopt the Common Core Standards. Most Texas districts still use the Professional Development Appraisal System (PDAS), which was developed in the mid 1990s; however, since 2011, some Texas districts have begun to explore options

for more updated evaluation systems that include student achievement as part of the assessment of teachers' work.

Evaluating teachers based on student achievement is common; however, districts vary in terms of the weight applied to each area of evaluation. A few examples include teacher practice, which is weighted at 90% for the teachers in the Chicago Public Schools and 40% for teachers in Florida. Test scores in Florida, however are 50% of a teacher's final rating, whereas students' test scores are not a factor for teachers in the Elgin Public School District.

Teacher evaluation across academic areas is an ever-changing landscape. It seems obvious that the structure of this enterprise has not been standardized, especially where non-tested subject areas are concerned. In the district where I conducted this study, for example, a music teacher's evaluation depends on the reading scores in her school. It remains unclear how school districts will continue to address the area of non-tested subjects, but considering the potential implications for teachers based on the results of their evaluations, it demands further examination.

Although music teacher evaluation is the focus of this project, all the teacher evaluation procedures that are discussed in this review are universally applied and used across all disciplines, as there are currently no specific evaluations that differ by subject area.

Teacher evaluations may be more efficient when descriptions of effective teaching are clarified and sequenced for an observer. Rehearsal frames have been applied in music observations (e.g., Cavitt, 2003; Colprit, 2000; Derby, 2001; Duke, 1994; Duke & Simmons, 2006; Montemayor, 2014; Worthy, 2003, 2006; Worthy & Thompson, 2009) and illustrate that music is an advantageous area in which to observe brief durations of teaching, as music classes include demonstrations of observable changes in

student performance, in the moment. In the following chapters, I describe my examination of the effects of different durations of observations on evaluations of music teaching.

Chapter Three: Method

The purpose of the present study was to determine the effect of observation duration on experienced evaluators' ratings of choral and instrumental music teaching. Ten evaluators from a large school district in Florida rated rehearsal frames and full rehearsals taught by 12 secondary-level choral and instrumental music teachers. The evaluators first observed and rated a video recording of two rehearsal frames excerpted from each teacher's full rehearsal and then observed and rated a recording of each teacher's full rehearsal. The evaluators rated the teachers on 10 criteria. I examined the evaluation scores to determine:

1. To what extent are evaluators' assessments of teaching affected by the duration of the teaching episodes they observe? Do ratings of teaching effectiveness differ between observations of brief, targeted excerpts and observations of full class periods?
2. To what extent do evaluators' levels of confidence in their assessments differ between these two observation conditions?

PARTICIPANTS

At the time of the study, the teacher participants (N = 12) were high school band and choral teachers in a large public school district in Florida (approximately 72,000 students enrolled in 85 schools). A description of their demographics appears in Table 3.1.

The 10 experienced evaluators (see Table 3.2) who participated in this study came from different academic backgrounds including English, math, history, physical education, music, and general education. The evaluators were principals (n = 8) and music supervisors (n = 2). All had had extensive experience evaluating music teachers as part of their professional responsibilities. As the former music supervisor of the district where the study was conducted, I knew the participants personally and had direct knowledge of their work.

Table 3.1
Teacher Demographics

								Years of Experience		
N	F	M	MS	HS	J/S-HS	Band	Choir	Range	*M*	*SD*
12	8	4	7	3	2	8	4	5-33	17.9	10.09

Note: Range of years of experience; *M* = Mean years of experience; *SD* = Standard Deviation; F = Female; M = Male; MS = Middle School; HS = High School; J/S-HS = Junior/Senior High School

Table 3.2 shows the demographics of the evaluator participants. It should be noted that the evaluators' years listed are the years they have been in a position where their duties included evaluating teachers (i.e., not overall years in education).

Table 3.2
Evaluator Demographics

							Years of Experience		
N	F	M	MS	HS	J/S-HS	Music Supervisors	Range	*M*	*SD*
10	5	5	5	2	1	2	4-26	15.5	6.70

Note: Range of years of experience; *M* = Mean years of experience; *SD* = Standard Deviation; F = Female; M = Male; MS = Middle School; HS = High School; J/S-HS = Junior/Senior High School

LOGISTICS AND PERMISSION

Permission was granted from the school district where the teacher and evaluator participants were employed and the protocol was approved by The University of Texas at Austin's Institutional Review Board (IRB). The documentation associated with these permissions appears in Appendices A and B.

I first contacted each of the prospective participants by phone, describing the nature of the investigation and their role in the study. All of the 12 teachers and 10 evaluators whom I contacted agreed to take part. After the evaluation phase of the study had begun, I replaced two of the principals who were unable to complete the evaluation tasks on schedule with two other principals who agreed to participate.

VIDEO DATA COLLECTION

Prior to my recording the rehearsals, I collected participant consent forms (see Appendices C and D) in accordance with the instructions provided by the Institutional Review Board of The University of Texas at Austin and the school district where the study was conducted. All consent forms were stored in a locked file cabinet in my office for the duration of the study.

Rehearsals were recorded in the teachers' home schools with their regular ensembles over a 6-day period in December of 2012. I recorded one rehearsal taught by each teacher, after inviting teachers to determine which of their ensembles they would like for me to record. Rehearsals were recorded using a Panasonic HD AVCCAM (model #AG-HMC40P) video camera mounted on a stationary tripod positioned at the rear of the classroom, focused on the teacher. Because of the district's student video policy, I was asked to focus the camera only on the teacher.

Directly after each rehearsal, I interviewed the teacher (see Appendix E for interview questions) about the demographics of the student population and information pertaining to the teacher's educational and professional experience.

Upon completion of each day of recording, I transferred the videos from the camera onto the hard drive of a Macbook Pro computer. Using iMovie software, I then compressed the original video files for ease of storage and editing. I labeled recordings with code numbers from a corresponding name-to-number log and stored backup copies of the files, without personal identifying information, on a Western Digital 1TB external hard drive.

VIDEO ANALYSIS OF REHEARSAL FRAMES

I began the analysis of each video by creating a timeline of the full rehearsal, noting the beginning and ending times of each activity. I then reviewed the videos to identify the rehearsal frames in each rehearsal. By definition, rehearsal frames begin with the implicit or explicit identification of a proximal performance goal (or target) and end when the target is successfully accomplished or abandoned. As expected, each rehearsal contained multiple successful and unsuccessful rehearsal frames.

To be certain that an evaluator would be able to assess the effectiveness of a teacher's instruction skills, I selected only rehearsal frames that included clear identifiable goals, and a clear indication whether the goal was accomplished or abandoned. I also made certain that each frame included multiple performance trials. As was observed in other research using rehearsal frames (e.g., Cavitt, 2003; Colprit, 2000; Worthy, 2006), there were several instances where positive changes in music

performance behavior required only a single verbal or non-verbal directive from the teacher, with the target accomplished during the next performance trial.

My goal was to choose rehearsal frames that provided what I thought was the best evidence of a teachers' capabilities. Fifty-seven frames (41 successful and 16 unsuccessful) were excerpted from the 12 rehearsal recordings; 4 to 6 frames from each rehearsal that most clearly depicted a teacher identifying a goal, then either accomplishing the goal or not.

Finally, I selected the two clearest successful rehearsal frames from each rehearsal to be evaluated in the first phase of the study, those in which the teacher explicitly or implicitly identified a target behavior and led the students to the accomplishment of the target. (There were times when the teacher did not explicitly state the target goal. For example, one teacher was attempting to correct a throat register A with a clarinet student and trying different approaches to get the A in tune. Although the teacher did not explicitly mention the word intonation, it was evident that intonation was the focus of the rehearsal frame.) The selected rehearsal frames ranged in duration from 58 seconds to 3.5 minutes.

As might be expected given the nature of typical school band or choir rehearsals, the unedited full-rehearsal videos included a variety of activities. In addition to rehearsing repertoire, full videos included warm-up activities, tuning, vowel exercises, sight-reading, and announcements. I edited the full videos to exclude announcements at the beginning or ending of class. The full videos ranged in duration from 40 to 55 minutes.

EVALUATION PROCEDURE AND CRITERIA

Instructions for completing the forms were e-mailed to evaluators (see Appendix F), and an online file-sharing service (Dropbox™) was used for exchanging the teaching videos and evaluation forms.

Evaluators received videos and evaluation forms (see Appendices G and H) for two teachers approximately once each week from February through April 2013. Evaluators were assigned a number from 1 through 10. Teachers were assigned a number from 1 through 12. Groups of participants were partially counterbalanced to control for order effects. Evaluators numbered 1-5 received two videos per week in the order of teachers 1-12; evaluators numbered 6-10 received two videos per week starting with teachers 7-12, followed by teachers 1-6.

Each evaluator viewed a total of 12 full-rehearsal videos and 24 rehearsal frames. Evaluators observed and rated the two rehearsal frames first, then observed and evaluated the full videos.

In all correspondence with the evaluators and on all of the written evaluation forms, the videos containing the two rehearsal frames were referred to as brief excerpts and the full rehearsals were referred to as full videos.

In selecting evaluation criteria for this study, I sought to create a list of 10 evaluation criteria that were consistent with the following goals: (1) the criteria needed to be commonly used across the selected districts outlined in Chapter 2; (2) the number of criteria needed to be manageable so as not to overburden the evaluators; and (3) the evaluators needed to be familiar with the language defining the criteria.

I examined the Florida districts' form where my sample of evaluators was employed and eliminated criteria from the form that were not in-class teacher behaviors. To create a manageable set of 10 in-class criteria from the criteria that remained, I

examined in detail 15 systematic evaluation systems used in the three most populous school districts in the five most populous states in the United States (http://www.census.gov, n.d.). The states and their districts are California (Los Angeles, San Diego, Long Beach); Texas (Houston, Dallas, Ft. Worth); Illinois (Chicago, Elgin, Rockford); New York (New York City, Buffalo, Rochester); and Florida (Dade, Broward, Hillsborough.) Chapter Two outlines a detailed list of each state and districts' criteria used for teacher evaluation.

Because many of the districts used the Danielson or Marzano evaluation systems, there were similar criteria that were prevalent throughout each of the 15 districts studied. I used the commonly addressed criteria (from the five states' criteria examined) of in-class teaching behaviors as a guide to select the 10 criteria from the Florida districts' form. The criteria for this study were found in all 15 districts' evaluation systems that were examined.

To make certain all evaluators were working from the same operational definitions for teacher behavior, I included a definition of each criterion on the back of the evaluation forms (see Appendix G). The criteria chosen for this study appear below.

1. Demonstrates evidence of planning and organization
2. Delivers engaging, challenging, and relevant lessons
3. Uses instructional time effectively
4. Demonstrates evidence of classroom management
5. Communicates to students clearly
6. Demonstrates knowledge of subject matter
7. Differentiates instruction
8. Provides instructional assessment

9. Identifies gaps in student's learning and modifies instruction in response to student misconceptions

10. Demonstrates knowledge of students

Using one evaluation form for the video with two rehearsal frames, and one form for the full video, evaluators rated each of the 10 criteria along 12-point scales. The 12 scale points were grouped into four categories labeled *Distinguished* (10-12), *Proficient* (7-9), *Needs Professional Support* (4-6), and *Unsatisfactory* (1-3). In addition to rating the teacher on each criterion, evaluators rated their own confidence level for each rating on a 5-point scale.

After viewing the full video for each teacher, evaluators also indicated how much of the video they thought they would have needed to view in order to provide an accurate assessment of the teacher's work. Their choices were: *100%, 75%, 50%, 25%*. Twenty-four evaluation forms per evaluator were collected and analyzed (12 rehearsal frames and 12 full video x 10 evaluators), totaling 240 forms.

POST HOC INTERVIEW QUESTIONS

I conducted post hoc interviews to gather more information about the evaluators' perceptions regarding the procedures of the study and their ideas about teacher evaluation in general. Questions are listed below with evaluators' answers provided in the following chapter.

1. What are your thoughts on teacher evaluation in general, and do you think the current protocol in your school district contributes to improvement in teaching, as well as improvements in student learning outcomes?

2. If you had an opportunity to change anything about the current protocol in your district, what would it be?

3. For teacher evaluations in general, in what ways can you imagine using video recordings in the formal evaluations of teachers, either as a sole source of data or as a complement to a live observation?

4. After evaluating the brief excerpts, how much and what kind of additional information do you think you gained after watching the full-length videos?

5. Given a hypothetical teacher evaluation procedure, please estimate the amount of information that several different observation options provide and how much each would contribute to formulating a valid and reliable assessment of a teacher's work. If two live in-class observations (Option 1) are assigned a score of 100 (entirely sufficient to formulate a valid and reliable assessment), and no observation (Option 5) is assigned a score of 0 (not at all sufficient), what scores would you assign to the other three options?

Option 1: Two live in-class observations conducted on two different days. (100)

Option 2: One live in-class observation.

Option 3: One video recording of a full-length class.

Option 4: Two purposefully selected brief recordings of a teacher making changes in a students' or class performance.

Option 5: No observation of teaching. (0)

6. What logistical or procedural problems did you encounter in completing the evaluation task? What could I have arranged differently to facilitate your work?

7. Were your scores on the 12-point scales affected by the classification level names above them (*Distinguished, Proficient, Needs Professional Support, Unsatisfactory*)?

Chapter Four: Results

Purpose and Overview of the Research

The primary purpose of this study was to determine whether experienced evaluators' assessments of teaching are affected by the durations of the teaching examples they observe. I compared the perceptions of 10 experienced evaluators who rated 12 music teachers based on video recordings of brief rehearsal frames and video recordings of full rehearsals. For each teacher, the evaluators first observed recordings of two rehearsal frames and then recordings of the full rehearsals from which the rehearsal frames were excerpted. The evaluators rated the teachers on 10 criteria following each recording. I examined the evaluation scores to answer the following questions:

1. To what extent are evaluators' assessments of teaching affected by the duration of the teaching episodes they observe? Do ratings of teaching effectiveness differ between observations of brief, targeted excerpts and observations of full class periods?
2. To what extent do evaluators' levels of confidence in their assessments differ between these two observation conditions?

Evaluators rated the teaching episodes on the following 10 criteria, which I described in Chapter 3:

1. Demonstrates evidence of planning and organization
2. Delivers engaging, challenging, and relevant lessons
3. Uses instructional time effectively
4. Demonstrates evidence of classroom management

5. Communicates to students clearly

6. Demonstrates knowledge of subject matter

7. Differentiates instruction

8. Provides instructional assessment

9. Identifies gaps in students' learning and modifies instruction in response to student misconceptions

10. Demonstrates knowledge of students

RESULTS AND ANALYSIS

Criteria

After examining the responses of the evaluators, I noticed that in the rehearsal frame condition one evaluator rated two of the criteria for every teacher with a rating of "1." When questioned about the scores, the evaluator stated that she thought she could not effectively evaluate those criteria because she did not explicitly observe them. Since her scores were outliers compared to the scores of the other evaluators, when calculating the mean scores for teachers, I replaced her scores for the two criteria with the mean values from the other evaluators.

Ten evaluators scored 12 teachers on 10 criteria. Table 4.1 shows the mean score, standard deviation, and difference scores for each of the 10 criteria between the experts' evaluations of the rehearsal frames (RF) and the full videos (FV) averaged across all teachers and evaluators.

On each evaluation form there was a 12-point scale across four classification levels (*Distinguished, Proficient, Needs Professional Support,* and *Unsatisfactory*). As

will be illustrated in this chapter, most of the evaluators (7 of 10) ignored this aspect of the evaluation form and concentrated only on the 12-point scale.

Table 4.1
Evaluator Rating Means and Standard Deviations for the 10 Evaluation Criteria

	12-point Scale					
	RF		FV		Δ	
Criterion	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
1	8.37	1.53	9.06	2.01	0.69	0.62
2	8.22	1.67	9.02	1.97	0.80	0.79
3	8.97	1.64	9.28	2.06	0.31	0.59
4	9.04	1.54	9.37	1.80	0.33	0.49
5	8.99	1.68	9.23	2.00	0.23	0.63
6	9.40	1.47	9.83	1.71	0.43	0.51
7	8.19	1.63	8.40	2.58	0.21	0.57
8	8.62	1.46	9.22	1.98	0.60	0.49
9	8.47	1.62	9.11	2.09	0.64	0.60
10	8.29	1.65	8.38	2.71	0.09	0.40

Note: Scores range from 1-12; RF = Rehearsal Frames; FV = Full Video; Δ = Score change from RF mean to FV mean; *SD* = Standard Deviation. Means and *SD's* were averaged across teachers first, then evaluators.

The standard deviations represent the variance among the evaluators' means. Evaluators' ratings of the full videos were generally higher than their ratings of the rehearsal frames for all criteria. It is important to note that the mean difference between the two observation conditions is less than one point along a 12-point scale for every criterion.

Evaluators rated their confidence levels for each criterion rating on 5-point scales. Table 4.2 shows the mean confidence ratings, standard deviations, and rating differences between the RF and FV evaluations for each of the 10 criteria, averaged across all

teachers and evaluators. The standard deviations represent the variance among the evaluators' means.

Table 4.2
Evaluator Confidence Level Means and Standard Deviations for the 10 Evaluation Criteria

	5-point Scale					
	RF		FV		Δ	
Criterion	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
1	2.93	0.87	3.80	0.81	0.88	0.32
2	3.04	0.98	3.94	0.77	0.90	0.34
3	3.20	1.04	4.13	0.64	0.93	0.35
4	3.37	0.93	4.08	0.69	0.74	0.40
5	3.48	0.91	4.18	0.61	0.71	0.31
6	3.33	1.04	4.01	0.74	0.68	0.30
7	2.88	1.06	3.88	0.76	0.99	0.37
8	3.00	0.87	4.04	0.65	1.04	0.28
9	3.03	1.04	4.04	0.63	1.02	0.38
10	2.85	1.00	3.83	0.76	0.98	0.27

Note: Confidence scores range from 1-5; RF = Rehearsal Frames; FV = Full Video; *SD* = Standard Deviation; Δ = Score change from RF mean and *SD* to FV mean and *SD*. Means and *SD's* were averaged across teachers first, then evaluators.

For each criterion, the mean confidence ratings are higher for evaluations of the FV than for evaluations of the RF. The mean confidence ratings for the RF range from 2.85 to 3.48, and the mean confidence ratings for the FV range from 3.80 to 4.18. The standard deviations in Table 4.2 indicate that the variability among the evaluators' confidence scores was higher for the RF than for the FV.

Relationships among the criteria

I examined the relationships among ratings for the 10 criteria in the evaluations of the FV recordings, and found little differentiation among the means. Evaluators tended to rate teachers higher on criteria that described observable instructional behavior that was immediately visible—*Demonstrates knowledge of subject matter* (*M* = 9.83), *Demonstrates evidence of classroom management* (*M* = 9.37), *Uses instructional time effectively* (*M* = 9.28), *Communicates to students clearly* (*M* = 9.23), *Provides instructional assessment* (*M* = 9.22), *Identifies gaps in student's learning and modifies instruction in response to student misconceptions* (*M* = 9.11), *Demonstrates evidence of planning and organization* (*M* = 9.06), and *Delivers engaging, challenging, and relevant lessons* (*M* = 9.02)—and rated teachers lower on criteria that were less clearly observable during the lessons—*Differentiates instruction* (*M* = 8.40), and *Demonstrates knowledge of students* (*M* = 8.38).

Table 4.3
Bivariate Correlation Matrix of the 10 Evaluation Criteria as Rated in the Full Video

	CR 1	CR 2	CR 3	CR 4	CR 5	CR 6	CR 7	CR 8	CR 9	CR 10
CR 1	1.000									
CR 2	0.972	1.000								
CR 3	0.854	0.875	1.000							
CR 4	0.801	0.826	0.830	1.000						
CR 5	0.882	0.944	0.888	0.878	1.000					
CR 6	0.930	0.906	0.890	0.905	0.891	1.000				
CR 7	0.853	0.880	0.773	0.789	0.886	0.771	1.000			
CR 8	0.800	0.855	0.892	0.867	0.833	0.812	0.837	1.000		
CR 9	0.824	0.892	0.916	0.880	0.964	0.851	0.873	0.874	1.000	
CR 10	0.692	0.786	0.755	0.886	0.913	0.809	0.698	0.717	0.861	1.000

Note: CR = Criterion.

Table 4.3 shows the bivariate correlations among the 10 criteria for the evaluations of the FV. These consistently high correlations indicate a generalized response set and little differentiation among the individual criteria.

Given the results of the correlation matrix and the high bivariate correlations between all pairs of criteria, I factor analyzed the 10 criteria mean scores for the FV to determine the extent to which these criteria could be reduced to a smaller number of assessment variables. It seemed reasonable to perform this analysis with the FV means, as they were based on the longer of the two observation conditions.

Table 4.4
Factor Loadings for the 10 Evaluation Criteria as Rated in the Full Video

Criterion Number	*Criterion*	*Factor Scores*
1	Demonstrates evidence of planning and organization	0.926
2	Delivers engaging, challenging, and relevant lessons	0.961
3	Uses instructional time effectively	0.932
4	Demonstrates evidence of classroom management	0.930
5	Communicates to students clearly	0.976
6	Demonstrates knowledge of subject matter	0.942
7	Differentiates instruction	0.898
8	Provides instructional assessment	0.911
9	Identifies gaps in student's learning and modifies instruction	0.960
10	Demonstrates knowledge of students	0.871

Note: All 10 evaluation criteria loaded onto one factor, which I labeled *Teacher Effectiveness*.

Table 4.4 illustrates how the FV factor analysis loaded all 10 criteria onto one factor, which I labeled *Teacher Effectiveness*. This factor explained 86.7% of the total variance for the entire set of variables.

Given the high bivariate correlations among the scores for the individual criteria and the results of the factor analysis, I performed all future analyses using a single score for each teacher. I created this score for each teacher by first calculating a mean of the scores given by the 10 evaluators for each criterion and then calculating a mean of the 10 criteria scores. All statistical comparisons in the remaining analyses are based on these overall Teacher Effectiveness scores.

Teachers

In this section I examine the differences among individual teachers' overall Teacher Effectiveness scores in the two observation conditions. Table 4.5 is in order of teacher rank (based on their full video Teacher Effectiveness score) and shows each teacher's mean and standard deviation across evaluators in both observation conditions. The means ranged from a low of 7.83 to a high of 10.79. The differences between the means of adjacent teachers are not large, and in fact, in some cases the means are quite close. Note that the rank order of the FV and RF are the same with the exception of Teachers 6 and 12.

The standard deviations in Table 4.5 express the variance among evaluators for each teacher. This illustrates a positive correlation between teachers' overall Teacher Effectiveness scores and the variation among evaluators.

The data presented in Table 4.5 again illustrate the increase in overall Teacher Effectiveness scores between the evaluations of the RF (M = 8.65, SD = 1.86) and evaluations of the FV (M = 9.09, SD = 2.14). A paired-samples t-test indicated that this difference was statistically significant, $t(11) = 4.03$, $p = .002$, $d = .22$. It is important to

note that this difference, although statistically significant, is approximately one half point on a 12-point scale.

Table 4.5
Means and Standard Deviations of Overall Teacher Effectiveness Scores for Each Teacher in the RF and FV Observation Conditions

	12-point Scale					
	RF		FV		Δ	
Rank based on Teacher FV Score	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
1	9.55	0.89	10.79	1.10	1.24	1.55
2	9.45	1.03	10.09	1.28	0.64	1.53
3	9.26	1.06	9.96	1.11	0.70	1.53
4	9.13	0.87	9.52	1.27	0.39	1.80
5	8.57	1.07	9.22	2.07	0.65	2.13
6	8.69	0.97	9.13	1.64	0.44	1.52
7	8.61	1.03	9.07	1.81	0.46	1.80
8	8.45	1.43	8.75	2.05	0.30	1.85
9	8.38	1.02	8.48	1.56	0.10	1.49
10	8.03	1.02	8.15	0.93	0.12	1.86
11	7.61	1.52	8.06	2.28	0.45	1.99
12	8.10	1.42	7.83	2.18	- 0.27	1.63

Note: Rank based on FV Mean; Scores range from 1-12; *SD* = Standard Deviation; RF = Rehearsal Frames; FV = Full Video; Teachers in order by FV mean score (highest to lowest); Δ = Score change from RF mean to FV mean.

Using the overall Teaching Effectiveness scores, I examined the data for evidence of effects attributable to evaluator gender, teacher gender, and rehearsal type (band or choir). I also examined the relationship between evaluators' mean ratings and their confidence levels. The results of these analyses are below.

Is there a relationship between the overall Teacher Effectiveness means and their standard deviations?

Table 4.5 shows the means and standard deviations for each teacher's overall Teaching Effectiveness score in the RF and the FV observation conditions. The standard deviations in this table represent the variation among evaluators' ratings of each teacher. The standard deviations in the RF condition were lower than the FV for all teachers except for the teacher ranked tenth.

Using Pearson's *r,* I found a moderate correlation between the overall Teacher Effectiveness scores and the corresponding standard deviations, $r(12) = -.563$, $p < .05$ (computed as a two-tailed test). The mean scores ranged from 7.83 to 10.79. This analysis illustrates a greater disparity among evaluators when teachers' scores were lower.

I also examined the relative reliability among evaluators for scores based on the FV and the RF. I conducted this analysis in two ways: by examining the standard deviations for each teachers' overall Teacher Effectiveness score, which represents the variation among evaluators, and by computing intraclass correlation coefficients for the overall Teacher Effectiveness scores in the RF and FV conditions. The standard deviations indicate that there was less variability in the FV scores than in the RF scores. Estimates of reliability among evaluators were moderately high in both observation conditions. The intraclass correlation in the RF condition was .78, $F(11, 99) = 4.53$, $p < .001$; the intraclass correlation in the FV condition was .79, $F(11, 99) = 4.77$, $p < .001$.

Are there scoring differences that are attributable to gender?

There were four male and eight female teachers in this study. Although three of the four males were ranked first, second, and third, the lowest (12^{th}) ranked teacher was also a male. Teachers who were ranked fourth through eleventh were female. I found no

significant relationship between teacher effectiveness and gender on the basis of the Mann Whitney U test, $U = 24$, $p = .174$.

Further, I sought to determine if there was a difference in the way male and female evaluators scored the teachers that was attributable to gender. There were five male and five female evaluators. To rank evaluators ratings from highest to lowest, I averaged the Teacher Effectiveness scores that evaluators provided for each teacher's full video. The range of the scores, on a 12-point scale, was 7.55 to 11.28. I found no significant relationship between teacher effectiveness and evaluator gender on the basis of the Mann Whitney U test, $U = 17$, $p = .347$.

Are there differences attributable to rehearsal type in the way evaluators score the teachers of instrumental and choral ensembles?

There were eight band and four choir teachers in this study. Instrumental teachers are ranked first, second, third, and eighth through twelfth; choral teachers were ranked fourth through seventh. I found no significant relationship between teacher effectiveness and rehearsal type on the basis of the Mann Whitney U test, $U = 20$, $p = .497$.

Is there a relationship between the evaluators' confidence scores and the overall Teacher Effectiveness scores?

I asked evaluators to rate their confidence level on a scale from 1 to 5 for each criterion for the RF and the FV. The data in Table 4.6 illustrate that the mean confidence level of the evaluators for each teacher increased overall from the ratings based on the RF to the FV. For the full video, evaluators' confidence scores were the highest for the first- (4.43) and second- (4.19) ranked teacher. The evaluators' mean confidence score for the eleventh- and twelfth-ranked teacher, however, were 4.07 and 3.95, respectively,

indicating that the evaluators were relatively confident in their assessments even of teachers they deemed unsatisfactory.

Table 4.6
Means and Standard Deviations for Evaluator Confidence Scores for Each Teacher in the RF and FV Observation Conditions

	5-point Scale					
	RF		FV		Δ	
Rank based on Teacher FV Score	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
1	3.51	0.72	4.43	0.63	0.92	1.06
2	3.39	0.69	4.19	0.39	0.80	0.98
3	3.10	0.77	3.97	0.68	0.87	0.99
4	3.20	0.74	4.01	0.57	0.81	0.91
5	3.05	0.59	3.96	0.48	0.91	1.02
6	2.87	0.91	3.88	0.43	1.01	1.18
7	3.02	0.79	3.88	0.42	0.86	1.04
8	3.07	0.69	3.94	0.49	0.87	0.94
9	2.94	0.86	3.86	0.47	0.92	1.11
10	2.94	1.15	3.77	0.46	0.83	1.14
11	3.26	0.93	4.07	0.52	0.81	1.20
12	2.97	0.53	3.95	0.46	0.98	0.88

Note: Rank based on FV Mean; Confidence means range from 1-5; *SD* = Standard Deviation; RF = Rehearsal Frames; FV = Full Video; Teachers in order by FV mean score (highest to lowest); Δ = Score change from RF mean to FV mean.

Using Pearson's *r*, I found a moderate correlation between the evaluators' FV mean confidence score and the overall Teacher Effectiveness score across all evaluators, $r(12) = -.698$, $p < .05$, (computed as a two-tailed test). The mean confidence score range was 3.77 to 4.43; mean teacher score range was 7.83 to 10.79. According to their overall Teacher Effectiveness scores, evaluators expressed greater confidence in evaluating what they perceived to be the most effective and least effective teachers in the sample, and

were less confident about the teachers whom they considered to be in the middle of the sample.

Is there a relationship between the Teacher Effectiveness scores and observation duration preferences?

Given the fact that the evaluators observed the entire video, I sought to determine how much of the full video they thought they needed to view before they felt confident in providing an accurate assessment of a teacher. The full video evaluation form provided a space for evaluators to answer this question with the choices *100%, 75%, 50%* or *25%*. Table 4.7 shows Teacher Effectiveness score means, ranks, and percentages across all evaluators. The numbers under each percentage indicate the number of evaluators who chose that percentage as the amount of time they felt they needed to view the full video before providing a confident score to a teacher.

Results indicate that evaluators thought they needed less time to evaluate the highest ranked teachers than they needed to evaluate the moderately and lower ranked teachers. The two highest ranked teachers yielded the lowest mean time needed for observation of 55% and 57.5%, followed by the sixth ranked teacher with a mean score of 60%. Teachers fourth and eighth had a score of 67.5%. The highest mean score was 82.5% for the seventh ranked teacher.

These data indicate that on average, evaluators were able to confidently assess a teacher more quickly when their assessment was positive, while it took slightly longer for evaluators to confidently assess teachers they deemed less effective.

Table 4.7 illustrates that most evaluators clustered around the 50% to 75% range, with no evaluator needing to watch the entire video for the teacher who was ranked first.

Three of the evaluators indicated they did not need to watch the entire video for *any* teacher before they felt ready to provide a confident score.

Table 4.7
Teacher Effectiveness Score Means, Ranks, and Observation Duration Preferences

Teacher FV Mean Score	*Rank*	*100 %*	*75%*	*50%*	*25%*	*Mean*
10.79	1	0	3	6	1	55.0
10.09	2	1	2	6	1	57.5
9.96	3	1	3	6	0	62.0
9.52	4	4	1	3	2	67.5
9.22	5	3	3	4	0	72.5
9.13	6	1	4	3	2	60.0
9.07	7	6	2	1	1	82.5
8.75	8	2	3	5	0	67.5
8.48	9	2	7	1	0	75.0
8.15	10	3	3	3	1	70.0
8.06	11	2	3	3	2	62.5
7.83	12	3	3	3	1	70.0

Note: Teacher = mean score for FV based on range from 1-12. Rank = order of teachers based on means of FV score (1= highest/12 = lowest). Mean = percentage (across evaluators) of the FV that evaluators felt they needed to view before confidently providing a rating. Numbers under percentages indicate number of evaluators who selected each category for each teacher.

Using Pearson's r, I found a moderate inverse correlation between the overall Teacher Effectiveness scores and the evaluators' mean observation duration preferences, $r(12) = -.523$, $p < .05$ (computed as a two-tailed test). These data suggest that when teacher mean scores were higher, the amount of time evaluators thought they needed to confidently assess a teacher's work was lower.

Evaluators

The means and standard deviations for evaluators are shown in Table 4.8. There is an increase in scores between the observations of the RF and FV for all evaluators except one (E4). The standard deviation scores between the RF and the FV decreased for six of the 10 evaluators (though by a very small margin), and all but one of the evaluators (E1) had standard deviation scores that were less than one point for both observation conditions.

Table 4.8
Evaluators' Mean Scores for the RF and FV Observation Conditions

	12-point Scale					
	RF		FV		Δ	
Evaluator	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
1	7.38	3.26	7.55	2.31	0.17	0.95
2	9.88	0.48	11.28	0.29	1.40	0.35
3	8.51	0.52	9.68	0.46	1.17	0.28
4	8.21	0.50	8.04	0.51	- 0.17	0.26
5	8.28	0.32	8.85	0.18	0.57	0.17
6	8.51	0.66	9.36	0.49	0.85	0.43
7	8.09	0.43	9.18	0.50	1.09	0.33
8	7.51	0.62	7.75	0.67	0.24	0.32
9	8.99	0.68	9.23	0.49	0.24	0.37
10	9.83	0.42	9.97	0.55	0.14	0.36

Note: Scores are across all teachers and all criteria; Criterion score means range from 1-12; RF = Rehearsal Frames; FV = Full Video; *SD* = Standard Deviation; Δ = Score change from RF evaluations to FV evaluations. In calculations of individual teachers' scores in previous tables and analyses, I used corrected data for Evaluator 1's scores for Criteria 7 and 10. Table 4.8 presents Evaluator 1's uncorrected data.

Based on a paired-samples *t*-test using data averaged across teachers and criteria, I found a statistically significant difference between the scores for the RF ($M = 8.52$, *SD*

= .85) and the FV (M = 9.09, SD = 1.12) observation conditions, $t(9) = 3.41$, $p = .008$, d = .57. All of the evaluators tended to rate teachers higher in the FV than in the RF condition and the range of difference scores varied from 0.14 to 1.40. For 5 of the 10 evaluators, the mean differences in averaged scores between the two conditions was less than a quarter point; the means for 3 of the 10 evaluators were as large as one point or larger. Although not inconsequential, these differences seem rather small, their statistical significance notwithstanding, given that evaluators watched approximately 10 times as much instructional time in the FV than they watched in the RF.

Table 4.9
Evaluators' Mean Confidence Scores for the RF and FV Observation Conditions

	5-point Scale					
	RF		FV		Δ	
Evaluator	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
1	3.03	0.11	3.08	< 0.01	0.05	0.11
2	2.72	0.55	4.38	0.28	1.67	0.55
3	2.27	0.23	3.91	0.13	1.64	0.15
4	3.81	0.29	4.18	0.17	0.37	0.19
5	4.08	0.28	4.48	0.24	0.40	0.12
6	2.44	0.42	4.38	0.21	1.93	0.38
7	2.77	0.16	3.96	0.17	1.19	0.20
8	3.72	0.22	3.92	0.21	0.20	0.08
9	3.64	0.46	3.77	0.36	0.13	0.19
10	2.63	0.48	3.88	0.55	1.25	0.31

Note: Scores are across all teachers and all criteria; Confidence score means range from 1-5; RF = Rehearsal Frames; FV = Full Video; *SD* = Standard Deviation; Δ = Score change from RF evaluations to FV evaluations.

Table 4.9 shows the evaluators' mean confidence scores and standard deviations across all teachers. Note that all evaluators' scores increase between the two observation

conditions. The standard deviations illustrate the differences among the evaluators, showing some felt more confident than others when providing their ratings in each of the two observation conditions.

CONCLUSIONS

In this study, with this sample of teachers and evaluators, overall results indicate that although there is a statistical difference between observing the rehearsal frames compared to the full video, the differences are small. The evaluators in this study rated teachers higher after watching the full video than after watching the brief excerpts in their mean scores, classification levels, and their confidence ratings.

On average, the evaluators needed to watch between 50% and 75% of a given teacher's video before providing what they determined to be a confident assessment of a teacher. Teacher quality, as determined by full video mean scores in this study, had a moderate impact on the duration that an evaluator thought was needed before providing a confident score.

POST HOC INTERVIEW QUESTIONS AND SUMMARIES

I conducted post hoc interviews with the evaluators to learn more about the perceptions of their experience with teacher evaluation and their experience with this study. Listed below are the questions and summaries of their answers.

1. *What are your thoughts on teacher evaluation in general, and do you think the current protocol in your school district contributes to improvement in teaching as well as improvements in student learning outcomes?* As expected, every evaluator thought teacher evaluation was necessary. Eight mentioned the difficulty in having 50% of a

teachers' evaluation rating based on student test scores for several reasons, including the "unfairness" to non-tested subject areas (e.g., Art, Music, Physical Education). Most importantly, the evaluators thought the non-randomization of student placement in classes (students are often placed in classes grouped by ability level) did not fairly represent the overall effectiveness of a teacher. Seven of the 10 indicated they thought teacher evaluations improved both teacher and student growth, and three thought that teacher evaluations did not have an impact on the improvement of teaching or student achievement.

2. *If you had an opportunity to change anything about the current protocol in your district, what would it be?* Eight of the 10 evaluators stated they would increase the number of points given to in-class teaching behaviors (it is currently 21/100). Two evaluators said they would like to have more control over how to evaluate their teachers in what they term an "inflexible" system. One said that he would like for experienced teachers (those with more than 5 years' experience) who are new to the district not be held to same data collection demands as teachers with fewer than 5 years. Nearly all of the evaluators (9/10) listed two items that needed immediate attention: clarification of the rubric for teacher expertise that is used in their district, and clarification of the scoring for Value Added Modeling.

3. *For teacher evaluations in general, in what ways can you imagine using video recordings in the formal evaluations of teachers, either as a sole source of data or as a complement to a live observation?* No evaluators stated they could see video recordings as a sole source of data, as all wanted to be able to observe teachers in their classrooms; however, all said they would like to see it as a complement to live observations.

4. *After evaluating the brief excerpts in this study, how much and what kind of additional information do you think you gained after watching the full-length videos?* Six of the 10 evaluators said they saw more of the same information after watching the full video. Four said that some of the information was different between the two observation conditions. One of the four said she thought the information on the full video was probably the same as on the rehearsal frames, but said she wanted to watch the longer version to be sure no aspect of teaching was overlooked.

5. *Given a hypothetical teacher evaluation procedure, please estimate the amount of information that you think several different observation options provide and how much each would contribute to formulating a valid and reliable assessment of a teacher's work. If two live in-class observations (Option 1) are assigned a score of 100 (entirely sufficient to formulate a valid and reliable assessment), and no observation (Option 5) is assigned a score of 0 (not at all sufficient), what scores would you assign to the other three options?*

> Option 1: Two live in-class observations conducted on two different days. (100)
> Option 2: One live in-class observation. *Evaluators mean score: 43*
> Option 3: One video recording of a full-length class. *Evaluators mean score: 52*
> Option 4: Two purposefully selected brief recordings of a teacher making changes in a students' or class performance. *Evaluators mean score: 51*
> Option 5: No observation of teaching. (0)

I averaged the scores the evaluators provided for Options 2, 3, and 4 (italicized above). I also asked evaluators if their scores would have changed if the evaluator and teacher

could view the video together (for Options 3 & 4). All evaluators stated they would have increased their scores for Options 3 and 4 given the option of viewing the video with their teacher.

6. *What logistical or procedural problems did you encounter in completing the evaluation task in this study? What could I have arranged differently to facilitate your work?* The limitation of not having the camera focused on students in addition to the teacher was mentioned by all of the respondents. Also, several evaluators indicated they would like to have a written narrative provided with the videos designating each teacher's goal/objective for that day's lesson.

7. *Were your scores on the 12-point scales affected by the classification level names above them* (Distinguished, Proficient, Needs Professional Support, Unsatisfactory)? Three evaluators said that yes, their ratings were, in fact, affected by the classification levels shown above the 12-point scale. The remaining evaluators indicated that the level names did not affect their decisions in scoring.

Chapter Five: Discussion

Teacher evaluation is unquestionably one of the most frequently discussed topics in education. National efforts to reform education have attempted to focus attention on high quality teaching and the relationship between teaching and student accomplishment, going so far as to offer financial incentives to states and school districts for working to improve their teacher workforce through new, more stringent teacher evaluation procedures.

Although recent teacher evaluation procedures have attempted to shift the focus of attention from only evaluating teacher behavior to assessing student accomplishment, in large measure evaluations still tend to center on the documentation of specific teacher behaviors that are believed to be associated with quality instruction.

When student accomplishment is taken into account as a component of teacher evaluation, accomplishment is most often defined in terms of scores on annually administered standardized tests, even though the connections between specific teacher behavior and student tests scores have yet to be clearly defined. Assessments of student progress are often far removed from the act of teaching itself, and of course there are innumerable variables that affect student progress in school, many of which exist quite apart from the behavior of teachers.

Music education provides important opportunities for teacher assessment because evaluators can observe, in the moment, teachers effecting productive changes in student behavior: reshaping an embouchure, refining tone production, or correcting a rhythm. It could be argued that a teacher who cannot successfully change student performance in the short term is unlikely to effectively change student behavior over the course of a school year, although this assertion has yet to undergo empirical scrutiny. Likewise,

demonstrations of effective behavior change in the short term may serve as indicators of teacher effectiveness over the long term. And, as the current study was designed to investigate, evaluating teaching effectiveness by observing brief excerpts of instruction may be a way to create an evaluation system for music teachers that will be meaningful, efficient, and ultimately improve teachers' effectiveness (Duke, 1994).

The purpose of the present study was to determine whether expert evaluators' assessments of teachers vary between observations of rehearsal frames that demonstrate effective behavior change and observations of full class sessions. Ten experienced evaluators rated 12 music teachers on 10 criteria. The evaluators first observed brief video recordings of two rehearsal frames of each teacher and then a recording of a full class period taught by the same teacher. The evaluators rated the teachers on all 10 criteria following each observation. I examined the evaluation scores to determine:

1. To what extent are evaluators' assessments of teaching affected by the duration of the teaching episodes they observe? Do ratings of teaching effectiveness differ between observations of brief, targeted excerpts and observations of full class periods?
2. To what extent do evaluators' levels of confidence in their assessments differ between these two observation conditions?

In the discussion that follows, I explain my interpretation of the data presented in Chapter 4 and possible implications for teacher evaluation. Overall, I found that evaluators in the present study tended to rate teachers more highly and expressed greater confidence in their ratings in the FV condition than in the RF condition. These differences are quite clear and statistically significant, and lead to my conclusion that observing brief video episodes of teaching does not lead to the same ratings of teacher

effectiveness as does observing video recordings of full class sessions. It is also true that the differences I observed between the two conditions were larger in terms of evaluator confidence than in terms of ratings of teacher effectiveness.

Evaluators tended to see the teachers in the present study more positively the longer they had to observe their teaching. It is important to note that this effect was not consistent across teachers or evaluators, however. Although all but one teacher was rated more highly overall in the FV condition than in the RF condition, the differences between the two conditions varied considerably among teachers and among evaluators.

Evaluators tended not to differentiate among criteria

Initial examination of the individual evaluators' scores revealed very little differentiation among the 10 criteria in the evaluations of each teacher. In other words, teachers who were viewed positively tended to be rated highly on all 10 criteria. A bivariate correlation matrix confirmed this observation. The correlations among the criteria ratings indicate that the assessment of teaching, like the assessment of other complex behavior, often is derived from general impressions of quality, rather than from highly differentiated assessments of individual components of behavior.

A factor analysis of the mean evaluation scores for the FV condition revealed, perhaps not surprisingly, that all 10 criteria loaded onto a single factor. All of this led to my decision to use a mean evaluation score for each teacher (i.e., the mean of the scores on the 10 criteria) in the remainder of my analyses. In this experiment, the individual criterion scores seemed to provide little meaningful information beyond the overall means. The discussion that follows is based on the analyses of the mean overall scores, which on the basis of the factor analysis I labeled *Teacher Effectiveness*.

Evaluators differentiated among teachers, although all teachers were rated in the upper half of the rating scale in both observation conditions.

The teachers who agreed to participate in this study were all effective, and the evaluators' narrow range of scores confirm that assessment, although the evaluators differentiated among the teachers. The differences between the highest and lowest Teacher Effectiveness mean scores were approximately 1.5 points in the RF condition and approximately 3 points in the FV condition. I analyzed the reliability among the evaluators by calculating intraclass correlation coefficients for the RF and FV Teacher Effectiveness scores, and found moderately high reliability among evaluators in both conditions: .78 in RF and .79 in FV.

Reliability is an important part of evaluation in any context. Inter-rater reliability in music evaluations of applied music performances (Bergee, 2003; Fiske, 1977) and full ensemble performances (Garman, 1991; Hash, 2012) have been studied, and evaluator training has been cited as a way to increase inter-rater reliability among evaluators in teacher evaluation (Berliner, 1988, 1989; Haefele, 1993; Hallinger, Heck, & Murphy, 2014; Papay, 2012; Wise, Darling-Hammond, McLaughlin, & Berstein, 1985). It is understandable that ensuring all evaluators are following the same procedures and have a clear understanding of those procedures strengthens reliability. Although there was no in-person training in the current study, I contacted the evaluators by phone and e-mail to explain the protocol as well as their role in the study. Additionally, I was available throughout the study for potential questions or concerns.

Evaluators rated teachers more highly in the FV condition than in the RF condition.

I compared the evaluations of the RF to the evaluations of the FV in terms of 12-point rating scales and found the scores were significantly higher after evaluators

watched the FV than after they watched only the RF. Evaluators who participated in this study typically evaluate teachers based on observations of full class periods, and my post hoc discussions with the evaluators revealed that nearly all of them wanted to see more of the teachers' work in order to provide teachers every opportunity to obtain a high rating.

Yet, although statistically significant, the ratings on the 12-point scale increased by 7%, an interesting result in light of the scoring rubrics that appear in most teacher evaluation instruments, nearly all of which employ much smaller scale ranges (e.g., 1-4 or 1-5 scales). Thus, it seems doubtful that the magnitudes of differences between individual teachers' scores that I observed between the two observation conditions on the 12-point scales would result in teachers' being rated differently on the evaluation systems that are used in most districts. Of course, this conjecture should be tested empirically.

This is an important aspect to consider when allocating time to teacher assessments, as it indicates that evaluators rate video recordings of brief episodes of effective teaching and video recordings of full class periods quite similarly. Note again that the brief episodes I used in this study were not random samples of teaching but were purposely-selected rehearsal frames illustrating the teachers effecting changes in student performance.

Although the teachers in this study were all considered to be effective, I ranked the teachers based on the mean score from their FV to get an order of overall quality. After ranking the FV, I noticed that the ranks for the RF closely resembled the rankings of the FV. With exception of Teachers 6 and 12, the remainder of the teachers were in the same rank order based on the RF score as they were based on the FV score. This provides further evidence that evaluations based on brief recordings that illustrate teachers making changes in student behavior and evaluations based on recordings of entire class sessions, are quite similar.

In examining the standard deviations for the Teacher Effectiveness mean scores for each teacher (i.e., the variation among the 10 evaluators' scores), I noticed that 11 of the 12 teachers obtained higher deviations in the FV condition than they obtained in the RF condition. This indicates that the agreement among evaluators was somewhat lower in the FV condition than in the RF condition. Given the fact that there was more to observe in the FV condition and the fact that the rehearsal frames were selected purposefully to illustrate teachers effecting change in student behavior, it seems understandable that the variation among evaluators' ratings would be higher in the FV condition than in the RF condition.

All evaluators expressed greater confidence in their evaluations in the FV condition than in the RF condition.

It is perhaps not surprising then, given that the evaluators in this study were used to evaluating teachers based on full-class observations, that their FV confidence scores were higher (by 29%) than their RF confidence scores. When I examined the variability among the evaluators, I found that the confidence scores for some evaluators were very similar in the two conditions. In fact, 5 of 10 evaluators' difference scores between the two conditions were less than .40.

When comparing the Teacher Effectiveness scores with the evaluators' confidence scores, I found a moderate correlation between overall teacher quality and the confidence scores of the evaluators. I thought, as I did with the perceived observation duration of the full videos, that there would be a high correlation between the ratings of the teachers and the evaluators' confidence scores. The scores showed that the first- and second-ranked teachers had the two highest confidence level means and that evaluators also tended to have higher confidence ratings when scoring teachers in the sample who

were rated least effective; however, evaluators' confidence levels were lowest for teachers who were rated in the middle of the sample.

Differences between the RF and FV conditions were much greater in terms of evaluators' confidence than in terms of teachers' ratings.

Even though confidence ratings were higher when evaluators observed longer teaching episodes than when they observed shorter episodes, assessments of Teacher Effectiveness were higher by approximately 7% in the FV condition. The availability of additional observation time did not lead to markedly different assessments of teachers' skills, but did lead to different confidence levels among the evaluators. Confidence ratings in the FV condition were 29% higher overall than confidence ratings in the RF condition. The variation among the scores indicates that there were some evaluators who felt more confident with the FV than with the RF and others who felt equally confident in both conditions.

Evaluator confidence and observation duration

Evaluators' confidence in their ratings became a large part of this study, and perhaps more important than I had first anticipated. Evaluators often get a sense of the quality of a teacher's work early on in an observation and sometimes make quick decisions based on very little information (Ambady, 2010; Ambady & Rosenthal, 1993).

I examined evaluators' ratings of confidence and their perceptions of how much of a full rehearsal they thought they needed to observe in order to make a reliable judgment. Given that each evaluator observed video recordings of full rehearsals, I asked them to indicate how much of the FV they believed they needed to observe in order to

make a confident assessment of the teachers' work. Their choices were: *100%, 75%, 50%* or *25%*. I also wanted to determine whether the quality of teaching (as determined by Teacher Effectiveness scores) was related to the amount of time evaluators thought they needed to observe before providing a confident score. I thought that perhaps the mean observation duration scores would be lower for the teachers whose Teacher Effectiveness scores were higher, illustrating that it may not take as long to make a decision about a teacher's work if the teacher is effective.

As the data analysis in Chapter 4 indicate, I found a moderate inverse relationship between the perceived observation duration preferences and Teacher Effectiveness scores; when the teacher mean scores on the FV were higher, evaluators indicated they needed less viewing time to make a confident assessment.

Recall that most evaluators stated that they could have provided a score after observing 50% to 75% of the full video for all teachers. Few evaluators indicated that they would need as little as 25% or as much as 100% of the full rehearsal in order to provide a reliable score. This indicates that evaluators thought they needed to view more than what they viewed in the rehearsal frames, but not as much as a full class session. This is understandable as evaluators' comments in the post hoc interviews indicate that, especially regarding the less skillful teachers, they wanted to see more so they could give teachers every chance to succeed.

Searching to eliminate other factors that may have affected the evaluators' scores, I compared teacher rankings (as determined by the Teacher Effectiveness scores) to gender. The gender category included two sets of comparisons: overall mean scores for female teachers compared to male teachers; and scores from female evaluators compared to male evaluators. I also sought to determine whether the scores for choral and

instrumental teachers were clustered around one another, perhaps showing that teachers of instrumental classes were scored differently than teachers of choir classes.

The general trend of the scores suggests no advantage to either gender nor to either type of ensemble. The evaluators likely evaluated the teachers in much the same manner they do in their own schools: by looking for the criteria presented to them and attempting to be fair and reliable when observing the shorter and longer durations of teaching.

POST HOC INTERVIEWS

I conducted post hoc interviews to gather information about the evaluators' experience with the study and their opinions about teacher evaluation. Some conversations were lengthy, as several of the evaluators felt free to express opinions, both positive and negative, concerning the recent changes in teacher evaluation.

1. When asked about their thoughts on teacher evaluation in general, all of the evaluators deemed it necessary, but indicated that teacher quality was difficult to measure accurately. Rubrics, points, and value-added scores seemed to be a frustrating combination to undertake for the evaluators in this study, yet most could not suggest a better way to measure teacher effectiveness. Evaluators expressed frustration at the seemingly endless number of demands from the state legislature and the lack of time available for attempting to satisfy the requirements.

2. When asked about what changes they would make, absent any restrictions, no evaluators wanted to completely eliminate the current protocol in their district. Several evaluators mentioned making changes to the rubrics by giving more points to in-class teaching behaviors. The current rubric allocates only 21 points for in-class behaviors out

of a possible 100 points for the entire evaluation. Evaluators expressed concern over the possibility that teachers whom they considered ineffective might be evaluated as *Distinguished* if they scored well on the remaining 79 points of the rubric.

Two evaluators wanted more flexibility and control over the evaluation process. This is understandable, as evaluators are expected to know their teachers and school community better than anyone, yet they are provided with a prescribed set of evaluation procedures that are to be used for all teachers, in all schools, in all grade levels, across all content areas.

Regarding the value-added measure scores, evaluators stated that it is difficult to explain to teachers how their final evaluation rating is connected to value-added measures when the formula is so complex.

In the district where I conducted this study, 50% of a teacher's final evaluation score is tied to his students' test scores. Most evaluators stated that attaching 50% of any judgment about the effectiveness of a teacher on one test score is problematic. In this district, one of the main issues for evaluators is that the value-added measure scores for non-tested subject areas are not connected to the content the teachers cover. Currently, not all subject areas are given a standardized test; therefore, the reading portion of standardized tests for each school is used as the score for all teachers in the school who are not in tested subjects, regardless of their content area. Thus, for teachers who do not teach a tested subject, 50% of their score is affected by children's accomplishments in reading.

3. When asked about using video recordings in the formal evaluation of teachers, no evaluators said they would be comfortable using them as the sole source of data for making an assessment, but it is interesting that all said video recordings would be a welcome addition to complement the evaluation. I suspect the evaluators' resistance to

using video as the sole source is a result of years of observing teachers in a classroom setting and relying, in part, on what they say is "the feel" of the class. The evaluators stated they would like the idea of using videos to support an evaluation because of the number of teachers they evaluate. They indicated they would likely use the videos as an opportunity to refresh their memory on what they observed in the classroom.

Several evaluators stated they would like to view videos in a collaborative setting with the teacher. They expressed the need for longer post-evaluation conferences to give themselves time to help teachers correct any deficiencies; they could see how viewing a video with each teacher would assist them in accomplishing that goal.

4. When asked how much and what kind of additional information they received after watching the FV versus the RF, 6 of 10 evaluators agreed they saw teaching that was similar to what they had observed in the RF. Evaluators stated that, in general, they thought they were seeing more of the same kind of teaching from the shorter to the longer versions. Why then, on the following question, did the evaluators say on average they would prefer watching a longer version of a teacher's work to a shorter one?

5. I asked evaluators to assign a score from 0 to 100 on their preference for Options 2, 3, and 4, with 0 and 100 already provided for Options 1 and 5:

Option 1: Two live in-class observations conducted on two different days. (100)

Option 2: One live in-class observation. *Evaluators mean score: 43*

Option 3: One video recording of a full-length class. *Evaluators mean score: 52*

Option 4: Two purposefully selected brief recordings of a teacher making changes in a student or class's performance. *Evaluators mean score: 51*

Option 5: No observation of teaching. (0)

Most evaluators preferred the FV option to the option with two brief recordings, but still preferred both video options to one live, in-class observation. When I asked them about their option choices on this question, they stated, as reported earlier, that with the video options they could review the teacher's work as often as necessary. In terms of video duration, I suspect there is a certain comfort for evaluators to observe a teacher for longer than 8 minutes, which is the approximate amount of time of both RFs combined. In addition, some evaluators stated their intent to have what they view as the necessary documentation to be able to defend their decisions and to give teachers the highest justifiable rating by allowing themselves as much time as possible.

6. I asked the evaluators if there were any logistical problems they may have encountered during the study. All evaluators agreed that the instructions were clear and the flexibility of watching the videos on their own time was helpful. The evaluators stated they have had required observation training that centered on evaluating teachers based on the accomplishments of students. Thus, watching for student engagement is a large part of how they assess teachers. Due to video policies concerning students, I was unable to show students in the view of the camera. As stated in Chapter 3, recognizing the limitations of this aspect of the study, the evaluators were able to get a clear idea of the level of engagement (as evidenced by teacher/student interaction with playing and singing) of the classes as a whole, but perhaps not for individual students. The evaluators' inability to observe students was a departure from how they typically evaluate teachers.

Several evaluators indicated they would have liked to have a clearer idea of the objective for the class they were observing. Evaluators in this district are typically provided with routine information prior to a class observation. Teachers convey to the evaluator the ability and grade levels of the students, as well as the class goals for a given day. Evaluators stated it would have been helpful to have a narrative provided during

both video conditions with pertinent information about the class (grade levels, ability level, goals).

7. Finally, I asked the evaluators if the classification level names above the 12-point scales (e.g., *Distinguished, Proficient*) affected their decisions when providing a score. Three evaluators said their ratings were affected by the names of the levels above the numbers. The remaining evaluators indicated that the category names did not affect their scores. This is an interesting finding, since evaluations are typically completed with the evaluator providing the final rating using a word (e.g., *Distinguished*) rather than using a number. On the other hand, it may be that, as some evaluators noted, they were thinking of the evaluation scale (1-12) as a Likert-type scale. The three evaluators who stated that they scored the teachers using the levels as a guide indicated that the words were helpful because they had a mental image of a distinguished teacher, and compared the teachers in this study to that model when scoring.

FREQUENT CHANGES IN EVALUATION PROCEDURES

As mentioned in the review of literature, the question of how best to evaluate teachers remains unsettled. This is evidenced by how teacher evaluations frequently change within districts. Districts follow state legislation for evaluations, but legislation often allows for flexibility and adaptations to fit the needs of individual school communities. For this reason, differences in teacher evaluation procedures among districts are common.

The frequency of change *within* districts, however, is often unsettling, and disrupts the entire enterprise of teacher evaluation. Perhaps the ongoing search for effective teacher evaluations is the result of a lack of agreement about what is important

in the behavior of teachers. The sense that there is so much to evaluate potentially paralyzes district and state leaders who are charged with not only the *establishment* of goals for teacher evaluations, but also the *accomplishment* of those goals.

SUMMARY

The primary goal of this study was to determine whether observing video recordings of RFs and video recordings of full rehearsals would result in similar evaluations of teaching.

The data in this study seem to refute a commonly held perception that evaluators require a full class period of observation time to formulate an assessment of a teacher's effectiveness. The similarity of evaluations in the RF and FV conditions in the present study indicates that observing instances of teachers effecting productive changes in student behavior may convey as much information as does observing recordings of full rehearsals.

Observing RFs efficiently highlights the connection between what teachers do and what learners accomplish. If evaluators observe examples of teachers changing the performance of students in the moment, they may have sufficient information to formulate accurate evaluations of teachers' effectiveness.

The participants in this study were not chosen randomly. I selected the evaluators and teachers based on my knowledge of their work and their willingness to participate. Although the teachers varied somewhat in terms of their effectiveness, they were all effective teachers. Any generalization to other teachers or evaluators outside of this sample of participants should be approached cautiously.

FUTURE RESEARCH

The concept of using RFs as a basis for teacher evaluation need not be limited to music instruction; however, because the results of daily music instruction are so clearly observable, it seems an advantageous area to begin refinement for the potential application to other academic domains.

To increase the accuracy and efficiency of teacher assessment, further research is needed to ensure that evaluators are not only seeing what they are *required* to observe (i.e., frequency of evaluations per year, number of minutes per evaluation), but also what is *important* to observe.

Appendices

APPENDIX A

University of Texas at Austin Institutional Review Board Consent Form

OFFICE OF RESEARCH SUPPORT

THE UNIVERSITY OF TEXAS AT AUSTIN

P.O. Box 7426, Austin, Texas 78713 · Mail Code A3200
(512) 471-8871 · FAX (512) 471-8873

FWA # 00002030

Date: 12/03/12

PI: Dalaine Chapman

Dept: Music

Title: Expert Evaluations of Teacher Quality in Brief and Extended Instrumental and Choral Music Teaching

Re: IRB Expedited Approval for Protocol Number 2012-11-0032

Dear Dalaine Chapman:

In accordance with the Federal Regulations the Institutional Review Board (IRB) reviewed the above referenced research study and found it met the requirements for approval under the Expedited category noted below for the following period of time: 11/30/2012 to 11/29/2013 . *Expires 12 a.m. [midnight] of this date.* If the research will be conducted at more than one site, you may initiate research at any site from which you have a letter granting you permission to conduct the research. You should retain a copy of the letter in your files.

Expedited category of approval:

☐ 1) Clinical studies of drugs and medical devices only when condition (a) or (b) is met. (a) Research on drugs for which an investigational new drug application (21 CFR Part 312) is not required. (Note: Research on marketed drugs that significantly increases the risks or decreases the acceptability of the risks associated with the use of the product is not eligible for expedited review). (b) Research on medical devices for which (i) an investigational device exemption application (21 CFR Part 812) is not required; or (ii) the medical device is cleared/approved for marketing and the medical device is being used in accordance with its cleared/approved labeling.

☐ 2) Collection of blood samples by finger stick, heel stick, ear stick, or venipuncture as follows: (a) from healthy, non-pregnant adults who weigh at least 110 pounds. For these subjects, the amounts drawn may not exceed 550 ml in an 8 week period and collection may not occur more frequently than 2 times per week; or (b) from other adults and children2, considering the age, weight, and health of the subjects, the collection procedure, the amount of blood to be collected, and the frequency with which it will be collected. For these subjects, the amount drawn may not exceed the lesser of 50 ml or 3 ml per kg in an 8 week period and collection may not occur more frequently than 2 times per week.

☐ 3) Prospective collection of biological specimens for research purposes by non-invasive means. Examples:
(a) Hair and nail clippings in a non-disfiguring manner.
(b) Deciduous teeth at time of exfoliation or if routine patient care indicates a need for extraction;
(c) Permanent teeth if routine patient care indicates a need for extraction.

(d) Excreta and external secretions (including sweat).
(e) Uncannulated saliva collected either in an un-stimulated fashion or stimulated by chewing gumbase or wax or by applying a dilute citric solution to the tongue.
(f) Placenta removed at delivery.
(g) Amniotic fluid obtained at the time of rupture of the membrane prior to or during labor.
(h) Supra- and subgingival dental plaque and calculus, provided the collection procedure is not more invasive than routine prophylactic scaling of the teeth and the process is accomplished in accordance with accepted prophylactic techniques.
(i) Mucosal and skin cells collected by buccal scraping or swab, skin swab, or mouth washings.
(j) Sputum collected after saline mist nebulization.

☐ 4) Collection of data through non-invasive procedures (not involving general anesthesia or sedation) routinely employed in clinical practice, excluding procedures involving x-rays or microwaves. Where medical devices are employed, they must be cleared/approved for marketing. (Studies intended to evaluate the safety and effectiveness of the medical device are not generally eligible for expedited review, including studies of cleared medical devices for new indications).
Examples:
(a) Physical sensors that are applied either to the surface of the body or at a distance and do not involve input of significant amounts of energy into the subject or an invasion of the subject's privacy.
(b) Weighing or testing sensory acuity.
(c) Magnetic resonance imaging.
(d) Electrocardiography, electroencephalography, thermography, detection of naturally occurring radioactivity, electroretinography, ultrasound, diagnostic infrared imaging, doppler blood flow, and echocardiography.
(e) Moderate exercise, muscular strength testing, body composition assessment, and flexibility testing where appropriate given the age, weight, and health of the individual.

☐ 5) Research involving materials (data, documents, records, or specimens) that have been collected, or will be collected solely for non-research purposes (such as medical treatment or diagnosis).
Note: Some research in this category may be exempt from the HHS regulations for the protection of human subjects. 45 CFR 46.101(b)(4). This listing refers only to research that is not exempt.

☒ 6) Collection of data from voice, video, digital, or image recordings made for research purposes.

☒ 7) Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.
Note: Some research in this category may be exempt from the HHS regulations for the protection of human subjects. 45 CFR 46.101(b)(2) and (b)(3). This listing refers only to research that is not exempt.

☒ Use the attached approved informed consent document(s).

☐ You have been granted a Waiver of Documentation of Consent according to 45 CFR 46.117 and/or 21 CFR 56.109(c)(1).

☐ You have been granted a Waiver of Informed Consent according to 45 CFR 46.116(d).

Responsibilities of the Principal Investigator:

1. Report immediately to the IRB any unanticipated problems.

2. Submit for review and approval by the IRB all modifications to the protocol or consent form(s). Ensure the proposed changes in the approved research are not applied without prior IRB review and approval, except when necessary to eliminate apparent immediate hazards to the subject. Changes in approved research implemented without IRB review and approval initiated to eliminate apparent immediate hazards to the subject must be promptly reported to the IRB, and will be reviewed under the unanticipated problems policy to determine whether the change was consistent with ensuring the subjects continued welfare.

3. Report any significant findings that become known in the course of the research that might affect the willingness of subjects to continue to participate.

4. Ensure that only persons formally approved by the IRB enroll subjects.

5. Use only a currently approved consent form, if applicable.
   Note: Approval periods are for 12 months or less.

6. Protect the confidentiality of all persons and personally identifiable data, and train your staff and collaborators on policies and procedures for ensuring the privacy and confidentiality of subjects and their information.

7. Submit a Continuing Review Application for continuing review by the IRB. Federal regulations require IRB review of on-going projects no less than once a year a reminder letter will be sent to you two months before your expiration date. If a reminder is not received from Office of Research Support (ORS) about your upcoming continuing review, it is still the primary responsibility of the Principal Investigator not to conduct research activities on or after the expiration date. The Continuing Review Application must be submitted, reviewed and approved, before the expiration date.

8. Upon completion of the research study, a Closure Report must be submitted to the ORS.

9. Include the IRB study number on all future correspondence relating to this protocol.

If you have any questions contact the ORS by phone at (512) 471-8871 or via e-mail at orsc@uts.cc.utexas.edu.

Sincerely,

James P. Wilson

James Wilson, Ph.D.
Institutional Review Board Chair

APPENDIX B

Brevard Public Schools Consent Forms

Office of Accountability, Testing, and Evaluations
Research Application

Briefly respond to the following questions. Your completed application should be no longer than three pages with the Assurances Form as an additional attachment.

Name of Applicant: Da-Laine Chapman

Title of research project:

Expert Evaluations of Teacher Quality in Brief and Extended Episodes of Instrumental and Choral Music Teaching

Date of Submission: November 7, 2012

Mailing Address:

Austin, Texas 78735

E-mail address: chapmad@aol.com

Phone Number:

Business/University Address: (Required for student research)

University of Texas at Austin
Butler School of Music
1 University Station E3100
Austin, TX 78712-0435

Faculty Sponsor/Phone: Dr. Robert A. Duke Ph: 512.471.0972

E-mail address: bobduke@austin.utexas.ed

2. Hypotheses of study:
In terms of determining teacher quality, what are the differences and similarities of viewing extended or brief episodes of teaching when evaluating a music teacher?

Brevard Public Schools

Office of Accountability, Testing, and Evaluations
Research Application

4. **Anticipated start date: December, 2012**
   Anticipated completion date: March, 2013

5. **Brevard County school(s) which will be involved: [redacted]**

6. **Briefly summarize your research design. Include instruments to be utilized and sources of data dependent on school/district records:**

The purpose of the present study is to compare evaluations of brief and extended recorded episodes of instrumental and choral music teaching. Twelve secondary music teachers of varying levels of expertise teaching an entire class session will be recorded teaching and entire class session, and from that video, I will identify three rehearsal frames of their teaching. Two of the rehearsal frames will culminate with the accomplishment of a target goal and one of the rehearsal frames will illustrate an unsuccessful attempt to accomplish a target goal.

Ten experienced evaluators (five principals and five music supervisors), will individually observe the teachers' three rehearsal frames and the full video. The evaluators will be asked to describe the teachers' behavior and to rate the teaching in the rehearsal frames and the full video using a rating system currently employed in their home school district: "Distinguished", "Proficient", "Professional Support Needed", or "Unsatisfactory." Evaluators will first evaluate the three rehearsal frames and then evaluate the recordings of the entire class sessions. My aim is to compare the extent to which viewing the full class session provides information relevant to the evaluation beyond that which is available in the rehearsal frames.

The impact on class activity for this project is minimal and involves recording one class of each of the 12 teachers. In addition, five principals from BPS will observe and evaluate the videos. The remaining 5 evaluators are music supervisors from other Florida districts.

Evaluators will use the current BPS summative evaluation form as the evaluation instrument.

Principals who agree to participate will not have teachers involved in the study. Names of teachers and schools will be kept confidential from all evaluators.

A brief questionnaire will be provided to the participatng teachers. The survey will ask questions such as name, e-mail address, phone number, years of teaching experience, and grade level/subject of the recorded class.

A video consent form for students in the participating teachers' class is attached.

7. **Will project involve [redacted]**

Brevard Public Schools

Office of Accountability, Testing, and Evaluations
Research Application

Assurances Form

I understand that I am requesting permission to engage in a research Project, and I am not requesting information pursuant to Open Records Legislation. If my research project requires participation with students, I understand that I may be subject to the appropriate School Board policy regarding background investigations, as well as any applicable costs associated. Additionally, if my request is granted, I agree to abide by all policies, rules and regulations of the District, INCLUDING THE SECURING OF WRITTEN PARENT PERMISSION PRIOR TO IMPLEMENTATION OF MY PROJECT.

x Da. Faire Chapman — 11/8/12

Researcher — Date

I have read the procedures for Research Projects in the Brevard County Public School System and understand that supervision of this project and responsibility for an outcome report rests with me. I also understand that the privileges of conducting future studies in the Brevard County Public School System is conditioned upon the fulfillment of such obligations.

x — 11/8/2012

Sponsor/Advisor of Research Project (signature required for student research) — Date

Approval of Office of Accountability, Testing and Evaluation*:

Victoria B. Wilkins — 11/9/12

Signature — Date

*Approval of the study at the district level does not obligate principals to participate in the proposed research.

Approval of Principal*:

Signature — Date

*The principal's signature suggests that the research project has been reviewed and that the school will participate, subject to the researcher's compliance with District policies.

4

CENTER FOR MUSIC LEARNING
THE UNIVERSITY OF TEXAS AT AUSTIN

School of Music • 1 University Station E3100 • Austin, Texas 78712-0435
(512) 471-2466 • FAX (512) 471-2467 • www.cml.music.utexas.edu

Dear Parent/Guardian,

My name is Da-Laine Chapman and I am a former Band/Orchestra director and Music Resource Teacher for Brevard Public schools. I am conducting a research project for partial fulfillment of the Doctor of Philosophy degree in music education. The purpose of this letter is to let you know that I will be videotaping your child's music teacher. There will be no interaction with students; however, in attempting to capture the teachers' work, the video may also capture some students as well. The video will be used for evaluation of the music teacher only.

This form serves as permission for your child to potentially appear on videotape.

Thank you,

Da-Laine Chapman
Doctoral Candidate
University of Texas at Austin

Student Name_____

Parent /Guardian Signature_____

APPENDIX C

Teacher Consent Form

IRB USE ONLY
Study Number:
Approval Date:
Expires:

Consent for Participation in Research

Title: Expert Evaluations of Teacher Quality in Brief and Extended Episodes of Instrumental and Choral Music Teaching

Introduction

The purpose of this form is to provide you information that may affect your decision as to whether or not to participate in this research study. The person performing the research will answer any of your questions. Read the information below and ask any questions you might have before deciding whether or not to take part. If you decide to be involved in this study, this form will be used to record your consent.

Purpose of the Study

You have been asked to participate in a research study about music teacher evaluation. The purpose of this study is to compare longer and shorter episodes of music teaching.

What will you to be asked to do?

If you agree to participate in this study, you will be asked to

- Complete a short informational survey
- Agree to allow the researcher to videotape you teaching one class, thus your participation **will** be **video** recorded.
- This study will take **one hour of your time**.

What are the risks involved in this study?

There are no foreseeable risks to participating in this study.

What are the possible benefits of this study?

The potential benefit for all participants is greater awareness of the possibilities for more efficient and meaningful music teacher evaluations.

Do you have to participate?

No, your participation is voluntary. You may decide not to participate at all or, if you start the study, you may withdraw at any time. Withdrawal or refusing to participate will not affect your relationship with The University of Texas at Austin or the researcher in any way.

Will there be any compensation?

You will not receive any type of payment for participating in this study.

What are my confidentiality or privacy protections when participating in this research study?

This study is confidential. Participants and their schools will be assigned a pseudonym for the duration of the study. Video recordings will be stored securely on the researchers' password protected computer and only the evaluators and research team will have access to the recordings, unless consent is given from the participant for use for educational purposes outside the research study, such as professional presentations or classroom demonstrations. The data resulting from your participation may be used for future research or be made

available to other researchers for research purposes not detailed within this consent form. The tapes will be retained for the purpose of retrospective analysis, if necessary.

Whom to contact with questions about the study?

Prior, during or after your participation you can contact the researcher **DaLaine Chapman** at

NOTE: Only include this statement if the study is Expedited or Full Board:
This study has been reviewed and approved by The University Institutional Review Board and the study number is **[STUDY NUMBER].**

Whom to contact with questions concerning your rights as a research participant?

For questions about your rights or any dissatisfaction with any part of this study, you can contact, anonymously if you wish, the Institutional Review Board by phone at (512) 471-8871 or email at orsc@uts.cc.utexas.edu.

Participation

If you agree to participate please sign the bottom of this form and return it to DaLaine Chapman either in person; or scanned and e-mailed , Austin, TX 78735.

Signature

You have been informed about this study's purpose, procedures, possible benefits and risks, and you have received a copy of this form. You have been given the opportunity to ask questions before you sign, and you have been told that you can ask other questions at any time. You voluntarily agree to participate in this study. By signing this form, you are not waiving any of your legal rights.

_____
Printed Name

_____ _____
Signature Date

As a representative of this study, I have explained the purpose, procedures, benefits, and the risks involved in this research study.

_____
Print Name of Person obtaining consent

_____ _____
Signature of Person obtaining consent Date

APPENDIX D

Evaluator Consent Form

IRB USE ONLY
Study Number: 2012-11-0032
Approval Date: 11/30/2012
Expires: 11/29/2013

Consent for Participation in Research

Expert Evaluations of Teacher Quality in Brief and Extended Episodes of Instrumental and Choral Music Teaching

Introduction

The purpose of this form is to provide you information that may affect your decision as to whether or not to participate in this research study. The person performing the research will answer any of your questions. Read the information below and ask any questions you might have before deciding whether or not to take part. If you decide to be involved in this study, this form will be used to record your consent.

Purpose of the Study

You have been asked to participate in a research study about music teacher evaluation. The purpose of this study is to compare longer and shorter episodes of music teaching.

What will you to be asked to do?

If you agree to participate in this study, you will be asked to

- Observe and evaluate 12 teachers' videotapes
- Fill out an evaluation form
- This study will take approximately 12 hours of your time.

What are the risks involved in this study?

There are no foreseeable risks to participating in this study.

What are the possible benefits of this study?

The potential benefit for all participants is greater awareness of the possibilities for more efficient and meaningful music teacher evaluations.

Do you have to participate?

No, your participation is voluntary. You may decide not to participate at all or, if you start the study, you may withdraw at any time. Withdrawal or refusing to participate will not affect your relationship with The University of Texas at Austin or the researcher in any way.

Will there be any compensation?

TBD

What are my confidentiality or privacy protections when participating in this research study?

This study is confidential. Participants and their schools will be assigned a pseudonym for the duration of the study. The data resulting from your participation may be used for future research or be made available to other researchers for research purposes not detailed within this consent form. The teachers' tapes will be retained for the purpose of retrospective analysis, if necessary.

Whom to contact with questions about the study?

Prior, during or after your participation you can contact the researcher **DaLaine Chapman** at [redacted] or send an email to [redacted]

This study has been reviewed and approved by The University Institutional Review Board and the study number is **2012-11-0032.**

Whom to contact with questions concerning your rights as a research participant?

For questions about your rights or any dissatisfaction with any part of this study, you can contact, anonymously if you wish, the Institutional Review Board by phone at (512) 471-8871 or email at orsc@uts.cc.utexas.edu.

Participation

Please sign the bottom of this form and return it to DaLaine Chapman either in person; or scanned and e-mailed [redacted] or mailed to [redacted] Austin, TX 78735.

Signature

You have been informed about this study's purpose, procedures, possible benefits and risks, and you have received a copy of this form. You have been given the opportunity to ask questions before you sign, and you have been told that you can ask other questions at any time. You voluntarily agree to participate in this study. By signing this form, you are not waiving any of your legal rights.

_____
Printed Name

_____ _____
Signature Date

As a representative of this study, I have explained the purpose, procedures, benefits, and the risks involved in this research study.

_____
Print Name of Person obtaining consent

_____ _____
Signature of Person obtaining consent Date

APPENDIX E

Teacher Information Form

Questions for Teachers

Name_____

E-Mail Address_____

School_____

Grade level of recorded class_____

Subject: Band Choir Music Supervisor

High School, Middle School, or Jr./Sr. High School

Years Teaching Experience_____

APPENDIX F

Participant Instructions

Instructions

Thank you for participating in this project. Two sets of videos (two separate teachers) will be placed in our shared folder in DropBox for evaluation each week. A 'set' for each teacher includes one long video and two brief videos of the same teacher. You will also have 2 evaluation forms per teacher.

Your task:

Evaluation Forms

Familiarize yourself with the evaluation forms that I've placed in our shared folder in DropBox. The only difference in the two forms is what they are named. There is one that will be used for your evaluation of the 2 "Brief Excerpts" video (yes—ONE form for both of the short videos combined), and one that will be used for the "Full Video". There are 10 teacher indicators on each form, and you are to give a rating to each indicator. The second page of each evaluation form has definitions for each of the 10 indicators.

Videos and labeling

Each teacher has 3 videos: 2 are very short and 1 is longer than the first two. The 2 short videos are edited clips taken from the full video that you will watch after the two shorter clips. Please watch *Teacher 1 Brief Excerpt #1* first (of the first teacher), followed immediately by *Teacher 1 Brief Excerpt #2* (first teacher) and fill out the evaluation form labeled "Video Observation Instrument-Brief Excerpts" (please be careful to watch the videos in the exact order that they are labeled). Save the 'Brief Excerpts' form as your initials then BE, followed by the teacher number. For example, if I were evaluating the brief excerpts for the first teacher, mine would be marked as "DCBE 1". DC (my initials) BE (Brief Excerpts) 1(first teacher).

Finally, you will watch *Teacher 1 Full Video* of the first teacher and fill out the evaluation form labeled "Video Observation Instrument- Full Video" and save the same way, but with one exception: you will use "FV" instead of "BE" (example: DCFV 1).

You will watch the next teachers' set using the same procedure as the first set. You may take breaks between a teachers' set of videos, but not between videos within each teachers' set.

Upon completion of both sets (2 teachers), please check to make certain that the completed evaluation forms and videos are in our shared folder in DropBox. E-mail me to let me know you are finished and I will remove them from our folder and replace them with 2 new sets (2 more teachers' videos and forms). We will do this weekly until all 12 teachers have been evaluated.

Please call me [redacted] should you have any questions. Thank you!

APPENDIX G

Brief Excerpts Evaluation Form and Glossary of Terms

Videotape Observation Instrument for BRIEF EXCERPTS

Click on one box to the right of each indicator below that best describes your evaluation of that indicator. Then click on a box in the far right column to indicate how confident are you in your response. ↓ INDICATORS ↓	Distinguished	Proficient	Needs Professional Support	Unsatisfactory	How confident are you in your response? 5-Extremely 4-Very 3-Moderately 2-Somewhat 1-Not very
1. Demonstrates evidence of planning and organization	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
2. Delivers engaging, challenging and relevant lessons	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
3. Uses instructional time effectively	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
4. Demonstrates evidence of classroom management	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
5. Communicates to students clearly	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
6. Demonstrates knowledge of subject matter	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
7. Differentiates instruction	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
8. Provides instructional assessment	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
9. Identifies gaps in students' learning and modifies instruction in response to student misconceptions	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
10. Demonstrates knowledge of students	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐

Definitions:

1) Shows evidence of planning and organization: Instructional goals and objectives are consistently clear and focus on student learning. Directions, procedures, and content are appropriate for and clear to all students. Plans routinely to provide for instruction to meet the needs of students with varied levels.

2) Delivers engaging, challenging and relevant lessons: High levels of rigor and relevance challenge students to be intellectually engaged throughout. Teacher clearly demonstrates and articulates how content relates and applies to instructional activities, life, work, and community.

3) Uses instructional time effectively: Lessons are designed to maximize productive time. Pace moves smoothly and efficiently as evidenced by transitions between activities. Learning experiences and activities are highly organized and efficiently facilitated by the teacher and students, each assuming responsibility for maximizing time for learning.

4) Shows evidence of classroom management: All students are engaged in learning experiences, discussions, questioning, and demonstrations of learning. Students are on-task and very few, if any, class disruptions occur. Classroom rules and standards of conduct upheld by all members of the classroom. Consequences for inappropriate behavior are reasonable, clear, and consistently applied. The teachers' monitoring of events in the classroom is subtle and proactive.

5) Communicates to students clearly: Directions, procedures, and feedback are clear to students and the teacher anticipates possible misunderstandings. Teacher's spoken language contains well-chosen vocabulary that enriches the lesson. Teacher finds opportunities to extend student vocabulary.

6) Demonstrates knowledge of subject matter: The teacher understands the central concepts, tools of inquiry, and structures of the discipline(s) and creates learning experiences that make these aspects of subject matter meaningful for students.

7) Differentiates instruction: Changes instruction efficiently and appropriately, addressing the unique learning differences of students. Instruction addresses the wide range of learning styles and abilities present in the classroom, allowing opportunities for success for each student.

8) Provides instructional assessment: Provides timely, meaningful, and consistent feedback during guided practice and discussions. Allows students opportunities to use feedback from instruction to improve their performance.

9) Identifies gaps in students' learning and modifies instruction in response to student misconceptions: The teacher uses a variety of checks for understanding during strategic points in the lesson to identify learning gaps and adjust instruction. The teacher anticipates problems and uses multiple intervention strategies to assist student understanding and performance.

10) Demonstrates knowledge of students: Teacher displays understanding of individual students' skill, knowledge, and language proficiency. Teacher possesses information about each student's learning and medical needs, collecting such information from a variety of sources. Teacher displays extensive understanding of how students learn and applies this knowledge to develop a positive relationship with individual students.

Appendix H

Full Video Evaluation Form

Videotape Observation Instrument for FULL VIDEO

Given that you observed the entire Full Video, approximately how much did you watch before you felt confident enough to provide your ratings?	All of it ☐	About 3/4 ☐	Half ☐	About 1/4 ☐

Click on one box to the right of each indicator below that best describes your evaluation of that indicator. Then click on a box in the far right column to demonstrate how confident are you in your response. ↓ INDICATORS ↓	Distinguished	Proficient	Needs Professional Support	Unsatisfactory	How confident are you in your response? 5-Extremely 4–Very 3-Moderately 2-Somewhat 1–Not very
1.Demonstrates evidence of planning and organization	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
2. Delivers engaging, challenging and relevant lessons	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
3. Uses instructional time effectively	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
4. Demonstrates evidence of classroom management	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
5. Communicates to students clearly	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
6. Demonstrates knowledge of subject matter	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
7. Differentiates instruction	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
8. Provides instructional assessment	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
9. Identifies gaps in students' learning and modifies instruction in response to student misconceptions	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐
10. Demonstrates knowledge of students	12 11 10 ☐ ☐ ☐	9 8 7 ☐ ☐ ☐	6 5 4 ☐ ☐ ☐	3 2 1 ☐ ☐ ☐	5 4 3 2 1 ☐ ☐ ☐ ☐ ☐

References

Ambady, N. (2010). The perils of pondering: Intuition and thin slice judgments. *Psychological Inquiry*, *21*(4), 271–278.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, *64*(3), 431.

American Recovery and Reinvestment Act of 2009, Pub. L. No. 111-5 123 Stat. 115, (2009).

Amrein-Beardsley, A., & Collins, C. (2012). The SAS education value-added assessment system (SAS® EVAAS®) in the Houston independent school district (HISD): Intended and unintended consequences. *Education Policy Analysis Archives*, *20*(0), 12.

Babad, E., Bernieri, F., & Rosenthal, R. (1991). Students as judges of teachers' verbal and nonverbal behavior. *American Educational Research Journal*, *28*(1), 211–234.

Bergee, M. J. (2003). Faculty interjudge reliability of music performance evaluation. *Journal of Research in Music Education*, *51*(2), 137–150.

Berliner, D. C. (1988). Implications of studies on expertise in pedagogy for teacher education and evaluation. *New Directions for Teacher Assessment*, 39–68.

Berliner, D. C. (1989). The place of process-product research in developing the agenda for research on teacher thinking. *Educational Psychologist*, *24*(4), 325–344. doi:10.1207/s15326985ep2404_1.

Bill and Melinda Gates Foundation. (n.d.). Retrieved from http://www.gatesfoundation.org/.

Brandt, C., Mathers, C., Oliva, M., Brown-Sims, M., & Hess, J. (2007). Examining district guidance to schools on teacher evaluation policies in the midwest region. Issues & answers. REL 2007-No. 030. *Regional Educational Laboratory Midwest*, 37.

Calandra, B., Brantley-Dias, L., Lee, J. K., & Fox, D. L. (2009). Using video editing to cultivate novice teachers' practice. *Journal of Research on Technology in Education*, *42*(1), 73–94.

California Department of Education. (n.d.) Retrieved from http://www.cde.ca.gov/.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement*, *73*(5), 757–783.

Cavitt, M. E. (2003). A descriptive analysis of error correction in instrumental music rehearsals. *Journal of Research in Music Education*, *51*(3), 218.

Charlotte Danielson. (n.d.) Retrieved from http://www.danielsongroup.org/.

Clements-Cortès, A. (2011). Designing an effective music teacher evaluation system (part one). *The Canadian Music Educator*, *53*(1), 13.

Coker, H., Medley, D. M., & Soar, R. S. (1980). How valid are expert opinions about effective teaching? *The Phi Delta Kappan*, *62*(2), 131–149.

Colby, S. A., Bradshaw, L. K., & Joyner, R. L. (2002). Teacher evaluation: A review of the literature. Retrieved from http://eric.ed.gov/.

Colprit, E. J. (2000). Observation and analysis of Suzuki string teaching. *Journal of Research in Music Education*, *48*(3), 206–221. doi:10.2307/3345394.

Common Core Standards Initiative. (n.d.). Retrieved from http://www.corestandards.org/.

Croft, M., Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D., & Whitehurst, G. J. (2011). Passing muster: evaluating teacher evaluation systems. *The Brookings Institution*. Retrieved October 20, 2013, from http://www.brookings.edu/.

Danielson, C. (1996). A framework for teaching. Retrieved from http://dvandkq.net/.

Danielson, C. (2001). New trends in teacher evaluation. *Educational Leadership*, *58*(5), 12–15.

Danielson, C. (2006). *Teacher Leadership that Strengthens Professional Practice*. ERIC. Retrieved from http://www.eric.ed.gov/.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching*. Association for Supervision & Curriculum Development. Retrieved from http://books.google.com/.

Danielson, C., & McGreal, T. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, *93*(6), 8–15.

Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research*, *53*(3), 285.

Derby, S. E. (2001). *Rehearsal of repertoire in elementary, middle, and high school choirs: How teachers effect change in student performance* (Ph.D.). The University of Texas at Austin, United States -- Texas. Retrieved from http://search.proquest.com.ezproxy.lib.utexas.edu/.

Duke, R.A., (1994). Bringing the art of rehearsing into focus: The rehearsal frame as a model for prescriptive analysis. *Journal of Band Research*, *30*(1), 78–95.

Duke R.A., (1999). Measures of instructional effectiveness in music research. *Bulletin of the Council for Research in Music Education*, *143*, 1–48.

Duke, R. A., & Buckner, J. (2009). Watching learners learn. *MTNA E-Journal*, *1*, 17–28.

Duke, R.A., & Simmons, A. L. (2006). The Nature of expertise: Narrative descriptions of 19 common elements observed in the lessons of three renowned artist-teachers. *Bulletin of the Council for Research in Music Education*, (170), 7–19.

Elementary and Secondary Education Act of 1965, Pub. L. No. 89-10 Stat. 79 (1965).

Ellett, C. D., & Garland, J. S. (1987). Teacher evaluation practices in our largest school districts: Are they measuring up to "state-of-the-art" systems? *Journal of Personnel Evaluation in Education*, *1*(1), 69–92.

Fagot, B., & Hagan, R. (1988). Is what we see what we get? Comparisons of taped and live observations. *Behavioral Assessment*, *10*(4), 367–374.

Fiske, H. E. (1977). Relationship of selected factors in trumpet performance adjudication reliability. *Journal of Research in Music Education*, *25*(4), 256–263.

Florida Department of Education. (n.d.) Retrieved from http://www.fldoe.org/.

Garman, B. R. (1991). Orchestra festival evaluations: Interjudge agreement and relationships between performance categories and final ratings. *Research Perspectives in Music Education*, *2*, 19–24.

Garrison, C., & Ehringhaus, M. (2007). Formative and summative assessments in the classroom. *National Middle School Association*. Retrieved from http://ccti.colfinder.org/.

Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. Viking Press. Retrieved from http://books.google.com/.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added*. Mathematica Policy Research. Retrieved from http://ideas.repec.org/.

Glickman, C. D., Gordon, S. P., & Ross-Gordon, J. M. (2004). *Supervision and instructional leadership: A developmental approach*. Allyn and Bacon.

Goals 2000: Educate America Act, Pub. L. No. 103-227, 108 Stat. 125 (1993).

Goldstein, J. (2003a). Making sense of distributed leadership: The case of peer assistance and review. *Educational Evaluation and Policy Analysis*, *25*(4), 397–421.

Goldstein, J. (2003b). *Teachers at the professional threshold: Distributing leadership responsibility for teacher evaluation*. Stanford University. Retrieved from http://en.scientificcommons.org/.

Goldstein, J. (2004). Making sense of distributed leadership: The case of peer assistance and review. *Educational Evaluation and Policy Analysis*, *26*(2), 173–197.

Goldstein, J. (2007). Easy to dance to: Solving the problems of teacher evaluation with peer assistance and review. *American Journal of Education*, *113*(3), 479–508.

Goldstein, J., & Noguera, P. A. (2006). A thoughtful approach to teacher evaluation. *Educational Leadership*, *63*(6), 31.

Haefele, D. L. (1993). Evaluating teachers: A call for change. *Journal of Personnel Evaluation in Education*, *7*(1), 21–31. doi:10.1007/BF00972346.

Hallinger, P., & Heck, R. H. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980-1995. *Educational Administration Quarterly*, *32*(1), 5-44.

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability*, 1–24.

Hartshorne, R., Heafner, T. L., & Petty, T. M. (2011). Evaluating modes of teacher preparation: A comparison of face to-face and remote observations of graduate interns. *Journal of Digital Learning in Teacher Education*, *27*(4), 154–164.

Hash, P. M. (2012). An analysis of the ratings and interrater reliability of high school band contests. *Journal of Research in Music Education*, *60*(1), 81–100.

Heath, R. W., & Nielson, M. A. (1974). The research basis for performance-based teacher education. *Review of Educational Research*, *44*(4), 463–484.

Hoff, D. J. (2004). Debate grows on true costs of school law. *Education Week*, *23*(21), 1.

Hogarth, R. M. (2001). *Educating intuition*. University of Chicago Press. Retrieved from http://books.google.com/.

Illinois State Board of Education. (n.d.) Retrieved from http://www.isbe.state.il.us/.

Improving America's Schools Act of 1994. Pub. L. No. 103-382, 108 Stat. 3906 (1993).

Jacob, B. A. (2007). The challenges of staffing urban schools with effective teachers. *The Future of Children*, *17*(1), 129–153.

Jellison, J. (In Press). *Including Everyone: Successful Music Learning for Children with Disabilities*. The University of Texas at Austin.

Kahneman, D. (2002). Maps of bounded rationality: A perspective on intuitive judgment and choice. *Nobel Prize Lecture*, *8*, 351–401.

Kahneman, D. (2013). A perspective on judgment and choice: Mapping bounded rationality. *Progress in Psychological Science around the World. Neural, Cognitive and Developmental Issues: Proceedings of the 28th International Congress of Psychology 1*, 1. Retrieved from http://books.google.com/.

Keller, B. (1998). Principal matters. *Education Week*, *18*(11), 25–27.

Keruskin, T. E. (2005). *The perceptions of high school principals on student achievement by conducting walkthroughs* (Ed.D.). University of Pittsburgh, United States -- Pennsylvania. Retrieved from http://search.proquest.com.ezproxy.lib.utexas.edu/.

Kimball, S. M., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly*, *45*(1), 34–70. doi:10.1177/0013161X08327549.

Koretz, D. (2008). A measured approach. *American Educator*, *32*(2), 18–39.

Krajewski, R. J. (1978). Secondary principals want to be instructional leaders. *The Phi Delta Kappan*, *60*(1), 65–65.

Kyriakides, L. (2005). Drawing from teacher effectiveness research and research into teacher interpersonal behaviour to establish a teacher evaluation system: A study on the use of student ratings to evaluate teacher behaviour. *Journal of Classroom Interaction*, *40*(2), 44–66.

Kyriakides, L., & Demetriou, D. (2007). Introducing a teacher evaluation system based on teacher effectiveness research: an investigation of stakeholders' perceptions. *Journal of Personnel Evaluation in Education*, *20*(1), 43–64.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47–67.

Loup, K. S., Garland, J. S., Ellett, C. D., & Rugutt, J. K. (1996). Ten years later: Findings from a replication of a study of teacher evaluation practices in our 100 largest school districts. *Journal of Personnel Evaluation in Education*, *10*(3), 203–226.

Marshall, K. (2005). It's time to rethink teacher supervision and evaluation. *Phi Delta Kappan*, *86*(10), 727–735.

Marzano, R. J. (1988). Dimensions of thinking: A framework for curriculum and instruction. ERIC. Retrieved from http://files.eric.ed.gov/

Marzano, R. J. (2007). Using action research and local models of instruction to enhance teaching. *Journal of Personnel Evaluation in Education*, *20*(3), 117–128.

Marzano, R. J., & Haystead, M. (2011). 2010-2011 Adams 50 instructional model study. Retrieved from http://www.sbsadams50.org/.

Marzano, R. J., Pickering, D., & McTighe, J. (1993). Assessing student outcomes: Performance assessment using the dimensions of learning model. ERIC. Retrieved from http://files.eric.ed.gov/.

Marzano, R. J., Pickering, D., & Pollock, J. E. (2001). Classroom instruction that works: Research-based strategies for increasing student achievement. Ascd. Retrieved from http://books.google.com/.

Massachusetts Educational Reform Act of 1993, Acts of 1993, §71-1-105 (1993).

Measures of Effective Teaching. (n.d.). Retrieved from http://www.metproject.org/.

Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *The Journal of Educational Research*, 242–247.

Mishra, P., & Koehler, M. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *The Teachers College Record*, *108*(6), 1017–1054.

Montemayor, M. (2014). Evaluative and behavioral correlates to intrarehearsal achievement in high school bands. *Journal of Research in Music Education*, *62*(1), 33–51.

National Board for Professional Teaching Standards. (n.d.). Retrieved from http://www.nbpts.org/.

National Center for Learning Disabilities. (n.d.). Retrieved from http://www.NCLD.org/

National Commission on Excellence in Education. (1983). A Nation at Risk: The imperative for educational reform. *The Elementary School Journal*, 84, (2), 112-130.

National Institute for Excellence in Teaching. (2012). Retrieved from http://www.niet.org/.

New York State Department of Education. (n.d.) Retrieved from http://www.nysed.gov/.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Noakes, L. A. (2009). Adapting the utilization-focused approach for teacher evaluation. *Journal of MultiDisciplinary Evaluation*, *6*(11), 83–88.

Ovando, M. N. (2001). Teachers' perceptions of a learner-centered teacher evaluation system. *Journal of Personnel Evaluation in Education*, *15*(3), 213–231.

Ovando, M. N., & Harris, B. M. (1993). Teachers' perceptions of the post-observation conference: Implications for formative evaluation. *Journal of Personnel Evaluation in Education*, *7*(4), 301–310.

Ovando, M. N., & Ramirez, A. (2007). Principals' instructional leadership within a teacher performance appraisal system: Enhancing students' academic success. *Journal of Personnel Evaluation in Education*, *20*(1), 85–110.

Palazuelos, A. E., & Conley, S. (2008). Choice in teacher evaluations. *Choice*. Retrieved from http://www.acsa.org/.

Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123–141.

Peterson, K. (2004a). Research on school teacher evaluation. *NASSP Bulletin*, *88*(639), 60–79.

Peterson, P.E. (2010). Supporters of Race to the Top outnumber opponents, but plurality of public has no opinion. *Education Next*, *10*(3). Retrieved from http://www.educationnext.org/.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common Core Standards, the new U.S. intended curriculum. *Educational Researcher*, *40*(3), 103–116. doi:10.3102/0013189X11405038.

Race to the Top Initiative. (n.d.). Retrieved from http://www2.ed.gov/.

Range, B. G., Scherz, S., Holt, C. R., & Young, S. (2011). Supervision and evaluation: The Wyoming perspective. *Educational Assessment, Evaluation and Accountability*, 1–23.

Robert Marzano. (n.d.) Retrieved from http://www.marzanoresearch.com/.

Roberts, N. K., & Hecht, J. B. (1996). VTLOGANL: Coding and analyzing videotaped data. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 76–82. doi:10.3758/BF03203639.

Rothstein, J. (2007). *Do Value-added Models Add Value? Tracking, Fixed Effects, and Causal Inference*. Center for Economic Policy Studies, Princeton University. Retrieved from http://www.uh.edu/.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537–571.

Scriven, M. (1988). Duty-based teacher evaluation. *Journal of Personnel Evaluation in Education*, *1*(4), 319–334.

Shapiro, S., & Spence, M. T. (1997). Managerial intuition: A conceptual and operational framework. *Business Horizons*, *40*(1), 63–68.

Shirley, D. A., & Langan-Fox, J. (1996). Intuition: A review of the literature. *Psychological Reports*, *79*(2), 563–584.

State Teacher Policy Yearbook National Summary Report. (2013). Retrieved from http://www.nctq.org/.

Stein, M. K., & Nelson, B. S. (2003). Leadership content knowledge. *Educational Evaluation and Policy Analysis*, *25*(4), 423–448. doi:10.3102/01623737025004423.

Stodolsky, S. S. (1984). Teacher evaluation: The limits of looking. *Educational Researcher*, *13*(9), 11–18.

Stronge, J. H. (2006). Teacher evaluation and school improvement: Improving the educational landscape. *Evaluating Teaching: A Guide to Current Thinking and Best Practice*, 1–23.

Sullivan, J. P. (2012). A Collaborative Effort: Peer Review and the History of Teacher Evaluations in Montgomery County, Maryland. *Harvard Educational Review*, *82*(1), 142–152.

Taylor, D. M. (2006). Refining learned repertoire for percussion instruments in an elementary setting. *Journal of Research in Music Education*, *54*(3), 231–243.

Teacher Advancement Program (n.d.) Retrieved from (http://www.tapsystem.org/).

Teachscape. (n.d.) Retrieved from http://www.teachscape.com/.

Texas Education Agency. (n.d.) Retrieved from http://www.tea.state.tx.us/.

Tienken, C. H., & Canton, D. (2009). National curriculum standards: Let's think it over. *AASA Journal of Scholarship and Practice*, *6*(3), 3–9.

Tuytens, M., & Devos, G. (2010). The influence of school leadership on teachers' perception of teacher evaluation policy. *Educational Studies*, *36*(5), 521–536. doi:10.1080/03055691003729054.

United States Census Bureau. (n.d.). Retrieved from http://www.census.gov/.

United States Department of Education. (n.d.). Retrieved from http://www.ed.gov/.

Veal, W. R., & MaKinster, J. G. (1999). Pedagogical content knowledge taxonomies. *Electronic Journal of Science Education*, *3*(4). Retrieved from http://ejse.southwestern.edu/.

West, R. E., Rich, P. J., Shepherd, C. E., Recesso, A., & Hannafin, M. J. (2009). Supporting induction teachers' development using performance-based video evidence. *Journal of Technology and Teacher Education*, *17*(3), 369–391.

Wise, A., Darling-Hammond, L., McLaughlin, M., Berstein,H. (1985). Teacher evaluation: A study of effective practices. *The Elementary School Journal*, *86*(1), 61–121.

Worthy, M. D. (2003). Rehearsal frame analysis of an expert wind conductor in high school vs. college band rehearsals. *Bulletin of the Council for Research in Music Education, 156*, 11–19.

Worthy, M. D. (2006). Observations of three expert wind conductors in college rehearsals. *Bulletin of the Council for Research in Music Education*, (168), 51–61.

Worthy, M. D., & Thompson, B. L. (2009). Observation and analysis of expert teaching in beginning band. *Bulletin of the Council for Research in Music Education*, *180*, 29–41.

Zepeda, S. J. (2006). Classroom-based assessments of teaching and learning. In J.H. Stronge (Ed.). *Evaluating Teaching: A guide to current thinking and best practice* (2nd ed., pp. 101–124). Corwin.

Vita

DaLaine Chapman is an active researcher and conductor/clinician having presented at numerous clinics and conferences nationwide. Her research interests include music teacher evaluation and assessment as well as the supervision of student teachers.

Her professional experience includes both teaching and administration. Prior to returning to graduate school, she was a music supervisor in Florida, working with 85 public schools and over 100 of some of the finest music teachers in Florida. Prior to that appointment she had extensive experience teaching band and orchestra at the secondary level.

Dr. Chapman holds Bachelors and Masters degrees in Music Education from The Florida State University, and a Ph.D. in Music and Human Learning from The University of Texas at Austin.

E-Mail address: dalainechapman@utexas.edu

This dissertation was typed by the author.